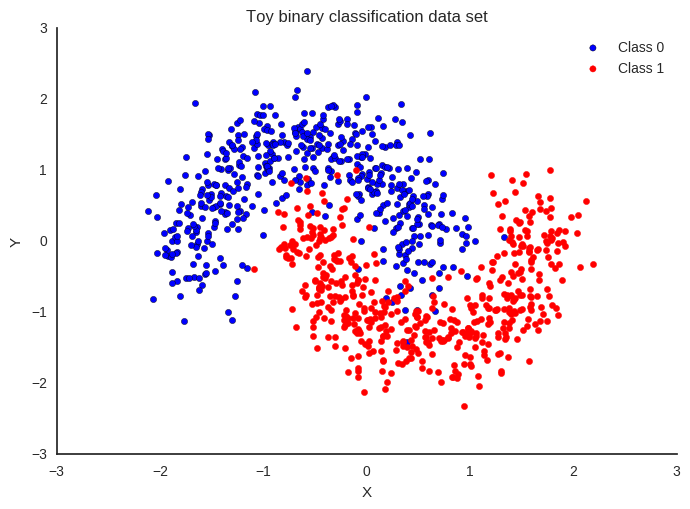


כריית ידע - פרויקט מסכם



מרצה: ד"ר אייל קולמן

עוזר הוראה: מר איתי מרגולין

מגישים:

|  |  |  |
| --- | --- | --- |
|  | שם | ת.ז. |
|  | נעם תור | 203302021 |
|  | תומר ינאי | 305187296 |
|  | יוגב מטלון | 201390408 |

תוכן עניינים

[תקציר 3](#_Toc485725263)

[מבוא 3](#_Toc485725264)

[מודל הסימולציה 4](#_Toc485725265)

[הנחות המודל 5](#_Toc485725266)

[תיאור המודל 5](#_Toc485725267)

[ניתוח המצב הקיים 7](#_Toc485725268)

[מדדים לבחינת המצב הקיים 7](#_Toc485725269)

[איכות השירות בפארק: 7](#_Toc485725270)

[מדדים תפעוליים: 7](#_Toc485725271)

[מדדי משרד הבריאות: 7](#_Toc485725272)

[סוג המערכת 7](#_Toc485725273)

[בחינת המדדים במודל הנוכחי 8](#_Toc485725274)

[הצעות לחלופה 10](#_Toc485725275)

[מבחן סטטיסטי השוואתי בין תוצאות מדדים 10](#_Toc485725276)

[מסקנות ודיון 11](#_Toc485725277)

[נספח א' – נתונים התחלתיים 12](#_Toc485725278)

[נספח ב' – נתונים אשר מקורם מחלק א' של הפרויקט 13](#_Toc485725279)

[נספח ג'- תיאור מפורט של חלקי המודל 14](#_Toc485725280)

[נספח ד' – חישובים נוספים 21](#_Toc485725281)

[חישוב מספר החזרות לסימולציה 21](#_Toc485725282)

[חישוב מדד סטטיסטי בין חלופות 21](#_Toc485725283)

# תקציר

TBD

# שלבי העבודה

שלב מספר 1 – Data Exploration

1. **בחירת נפוצים**– מבין תשעת הקלאסים הקיימים בדאטה בחרנו לסווג את המוצרים לשתי הקטגוריות הנפוצות ביותר: קטגוריה 2 ו-6. בעקבות בחירת קטגוריות אלו נותרנו עם 22,719 רשומות (מתוך 46,408).
2. **בדיקה האם התגיות מאוזנות** – האיזון חשוב מאוד כדי לוודא שהמודלים לא מסיקים מסקנות שגויות על סמך תגיות שהם לא מאוזנות (Data Label Distribution במחברת). ניתן לראות כי הלייבלים אכן מאוזנים (53% לקטגוריה 2 והשאר של 6).

**"מבט על" עבור הפיצ'רים** – ביצענו בדיקה עבור סוג (Type) כל פי'צר ובדיקה האם ישנם ערכים שאינם מספריים (או ערכי Null כמובן), תיאור סטטיסטי של כל אחד מהפיצ'רים (ממוצע, סטיית תקן, מינימום, מקסימום ורבעונים), ותרשים Box Plot לכל פי'צר. מניתוח ה-Box Plot ניתן לראות מעיכה דרמטית של הנתונים סביב ה-0 עבור כל הפיצ'רים.

1. **בדיקת התנהגות (התפלגות) הנתונים** – ביצענו בדיקות של התפלגויות הנתונים השונים (עבור כל feature) באמצעות גרפים העוזרים להבין את התנהגות הנתון בהשוואה לכל קטגוריה. בדיקות אלו מראות כי בכל אחד מהפיצ'רים הערך 0 הוא הנפוץ ביותר ובנוסף יכולות להעיד על קורלציה כפי שנפרט בהמשך. בנוסף ההתפלגויות יכולות לעזור לנו להבין אילו פיצ'רים יהיו הכי יעילים באבחנה בין שני הקלאסים- הפיצ'רים בהם השוני הגדול ביותר בין ההתפלגויות.

במקרה זה ההתפלגויות דומות יחסית.

בדקנו עבור כל פי'צר את מידת ה"ספארסיות" שלו – מה אחוז המדויק של הערכים שאינם אפס בכל פיצ'ר, ניתן לראות כי בפיצ'ר70 כמות המידע הגדולה ביותר ואילו פיצ'ר51 כמעט לא כולל רשומות שערכם אינו 0

1. **קורלציה** – בחלק זה בדקנו האם ישנה קורלציה בין הפיצ'רים השונים. קורלציה גבוהה תעיד כי עלינו לשקול הוצאת פיצ'רים, שכן אינן מסבירות מידע חדש. לאחר הבדיקה ראינו כי הפיצ'רים אינם קורלטיביים זה לזה, הקורלציות שנמצאו נמוכות מ0.4

שלב מספר 2 – Pre-Processing

1. **הסרת חריגים** – לא ביצענו הסרת חריגים מהנתונים. הסיבה לכך היא שמרבית הנתונים מאופסים כך שכפי שראינו בשלב הData exploration אפילו האחוזון ה75% בכלל הפיצ'רים הינו 0. כלומר הסרת חריגים תוריד רשומות חשובות בהם קיים מידע שאיננו 0, אך בנתונים אלו כל מידע שאיננו מאופס יקר ולכן חשוב לשמור רשומות אלו.
2. **Feature Extraction**- בחרנו להוסיף 2 עמודות לנתונים, בחרנו להוסיף 2 עמודות בלבד משום שלדעתנו עצם הספארסיות של הנתונים גורמת לכך שמניפולציות מתמטיות על הנתונים הקיימים לא יביאו הרבה תועלת- זאת משום שעיקר המידע כבר קיים בהאם נתון מסויים הוא 0 או לא, לכן לוגריתמים ופולינומים למיניהם יביאו תועלת מועטה.

עמודה “zeros\_number” – מכילה את כמות ה0 הקיימת ברשומה, לדוגמא רשומה בה כל הפיצ'רים הינם 0 תכיל את המספר 10 בעמודת ה’zeros\_number”.

הרעיון מאחורי עמודה זאת הוא לסכם בפיצ'ר אחד את כמות הפיצ'רים שאינם מאופסים ברשומה, זאת משום שלהערכתנו זהו מידע חשוב בסיווג הקלאסים.

עמודה “smart\_multy”- מכילה הכפלה של כלל המספרים ברשומה אחד בשני, אך משום שבכל רשומה קיים 0 אחד לפחות, הוספנו החרגה שהכפלה ב0 למעשה מהווה הכפלה ב1.

לדוגמא : רשומה המכילה בפיצ'ר מסויים 5 בפיצ'ר אחר 6 והשאר אפסים תקבל את הערך 30.

רשומה המכילה 10 אפסים (כלל הפיצ'רים) תקבל את הערך 1.

הרעיון מאחורי עמודה זאת הוא לתת משמעות חזקה ולא לינארית לכמות ולערך המספרים שאינם 0 ברשומה.

1. **נרמול הנתונים-** השתמשנו בשני סוגי נרמול ושמרנו את גם את הפיצ'רים המקוריים וגם את הפיצ'רים המנורמלים על מנת להחליט בהמשך באיזה סוג נרמול, אם בכלל, ברצוננו להשתמש.

נרמול MinMax-. הנרמול ממיר את הנתונים לאותה הסקלה – [0,1]. הנרמול מזיז את תוחלת ההתפלגות המקורית אך שומר על צורתה.

נרמול Z – הנרמול ממיר את הנתונים לסקאלה של "סטיות תקן", סקאלה שאיננה חסומה. כלל הנתונים מנורמלים לתוחלת 0 וסטיית תקן 1. היתרון של נרמול זה הוא שהוא מתאים יותר לשימוש בpca משום שסטיות התקן והתוחלות שוות.

1. **השלמת מידע חסר-** כפי שניתן לראות בסיכום הנתונים, לא קיימים נתונים חסרים בדאטה. וכפי שציינו נתייחס ל-0 כנתון בעל משמעות ולא כמידע חסר.
2. **הסרת עמודות-** לא הסרנו עמודות משום שאף על פי שיש עמודות בהם קיימים מעט מאוד ערכים, כל מידע חשוב בשל הספארסיות של הנתונים. וגם המעט מידע הקיים בעמודות אלו עלול להיות חשוב.
3. **שינוי type** – לצרכי נוחות החלפנו את עמודת הtarget לעמודה בינארית כאשר 1 מייצג שהרשומה שייכת לclass\_2 ו0 מייצג שהרשומה איננה שייכת לclass\_2, כלומר שייכת לclass\_6

בנוסף הפרדנו עמודה זאת לווקטור נפרד בשם Y, ולעמודות הפיצ'רים נתנו את השם X.

*שלב מספר 3 – Feature Selection*

ביצענו 3 בדיקות על מנת להבין האם ביכולתנו להוריד את מימד הנתונים וכיצד לעשות זאת:

**PCA-** את הpca ביצענו על העמודות המנורמלות בשיטת Z, זאת על מנת ששוני אפשרי בסקאלות של הפיצ'רים לא ישפיע על הקומפוננטות.

בשירטוט של שתי הקומפוננטות הראשונות ניתן לראות שלא נוצרת הפרדה טובה בין המחלקות אותם ברצוננו לסווג- נקודות שתי המחלקות חופפות אחת לשניה.

בבחינת הקומפוננטות שנוצרו אנו רואים שגרף מצטבר של אחוז השונות המוסברת עולה בצורה איטית ונזדקק לכמעט 10 קומפוננטות על מנת להסביר 95% מהשונות בנתונים. זאת כאשר המימד המקורי הינו 12. זוהי איננה הורדת מימד משמעותית ולכן לא נשתמש בpca להוריד מימד בבעיה.

**LDA-** את ההתאמה של קומפוננטות הlda ביצענו על העמודות המנורמלות בשיטת MinMax, אך בגרף שנוצר ניתן לראות שהקומפוננטה שנוצרה לא יוצרת הפרדה טובה בין המחלקות- יש חפיפה משמעותית בין הנקודות.

לכן הורדת מימד בשיטת LDA איננה מתאימה במקרה זה.

**Decision tree-** ביצענו אימון של עץ החלטה על הנתונים המנורמלים בשיטת MinMax, על מנת לראות מה הפיצ'רים שהעץ יבחר כמשמעותיים להפרדה בין הקלאסים. אימנו את העץ כאשר העומק המקסימלי הוא 24 (פעמיים מספר הפיצ'רים) והמדד הנבדק הוא gini. גם בשיטה זאת ניתן לראות שאם ברצוננו להגיע למעל 95% במדד הfeature importance של העץ עלינו לבחור עשרה פיצ'רים. הורדת מימד לא משמעותית.

בסיכום שיטות Feature selection בהם השתמשנו ניתן להגיד כי לא מצאנו שיטה טובה להורדת מימד משמעותי בנתונים תוך כדי שמירת המידע הכלול בהם. לכן הגדרנו את סט הtrain default להיות סט הנתונים הכולל את כל 12 הפיצ'רים.

למרות זאת אנו רוצים להשאיר לעצמנו בהמשך אפשרות לבדוק ביצועים של מודלים על הנתונים במימד נמוך יותר.

נבחר את הפיצ'רים שהיו משמעותיים בשיטת "עץ ההחלטה" ונגדיר בעזרתם דאטה-פריים חדש בו יכללו רק 5 מימדים- “feat\_61”, “feat\_20”, “feat\_70”, “feat\_51”, “smart\_multi”

כלל המימדים מנורמלים בשיטת MinMax, סכום הפיצ'רים במדד הfeature importance של העץ המאומן הוא 82.55%

*שלב מספר 4 – בניית מודלים "פשוטים"*

כשלב מקדים לשלב אימון המודלים ביצענו 2 פעולות:

1. חלוקת הנתונים בשיטת K-fold כאשר k=5, ושמירת הסטים המחולקים במערך לשימוש עתידי בשיטת cross validation בזמן אימון המודלים.
2. כתיבת פונקציה להצגת גרף ROC עבור תחזיות של מודל מסויים, חשוב לציין שהפונקציה נלקחה מלואה מתוך הדוקומנטציה של sklearn, רפרנס קיים בתוך מחברת הפרוייקט.

**KNN**

אימנו מודל KNN על סט הנתונים train\_defualt, ביצענו חיפוש בעזרת K-fold cross validation אחר הk המתאים ביותר לסט הנתונים שלנו.

המדדים שנבדקו על מנת להשוות בין הk השונים הינם “accuracy” ו"AUC" והחיפוש בוצע על K שגדלים בצורה אקפוננציאלית בחזקות של 2- כלומר K = 2,4,16…

ניתן לראות שכבר בk=8 המדדים עולים בצורה חדה ולאחר מכן מתקיימות עליות מינוריות יותר עד לk=64. בבחירת הk האופטימלי התלבטנו בין 64 ל32 משום שבk=64 מדד ה”AUC” גבוה יותר אך מדד ה"accuracy" נמוך יותר.

לבסוף החלטנו לבחור בk=32 משום שלמרות שההבדלים מאוד מינוריים הם טיפה משמעותיים יותר במדד ה”accuracy"

בk שנבחר המדדים שחושבו בשיטת הk-fold הם:

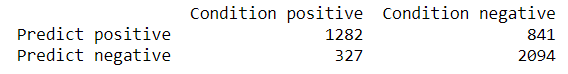
AUC = 0.77569

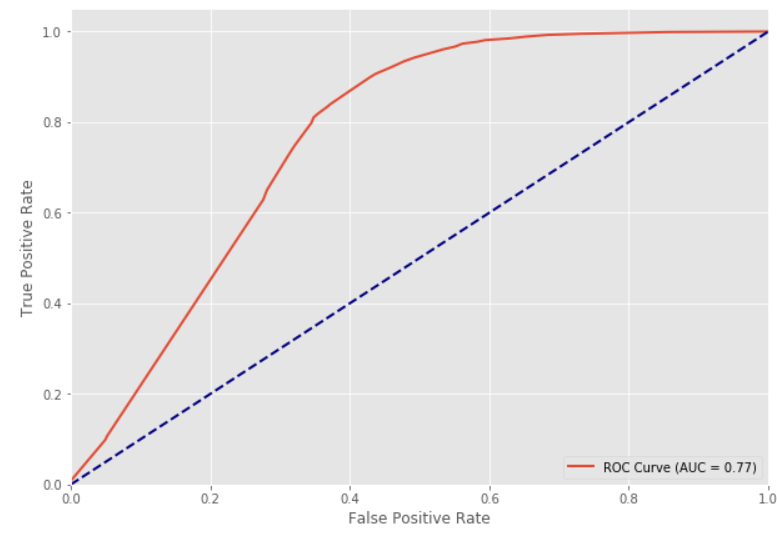
Accuracy = 0.74052

יש לציין שהמדדים חושבו בשיטת K-fold וכאשר נאמן את המודל על 100% מהנתונים הם עשויים להשתפר.

על מנת להציג גרף ROC ואת הConfusion matrix בחרנו fold רנדומאלי אימנו את המודל על סט הtrain שלו ובחנו על סט הtest שלו.

תוצאות הרצה זאת הינם:





**Logistic Regression**-

אימנו מודל רגרסיה לוגיסטית על סט הנתונים train default במלואו.

TBD

*שלב מספר 5 – בניית מודלים מתקדמים*

בשלב זה אימנו שני סוגי מודלים – Random Forest ורשת נוירונים.

משום שבשני המודלים יש מספר היפר-פרמטרים רב יחסית אשר יש להתאים לסוג הנתונים נעזרנו במחלקה מתוך sklearn בשם GridSearchCV אשר מריצה מספר רב של פעמים את המודל הנבחר, בכל פעם בשילוב אחר של פרמטרים אפשריים שהוגדרו. כל הרצה עובדת בשיטת k-fold בערך דיפולטיבי של k=3. מתודות המחלקה מאפשרות למצוא את הפרמטרים המביאים לתוצאה הטובה ביותר של המדד שהוגדר, ושומרות מודל מאומן בפרמטרים אלו בו ניתן להשתמש.

**Random Forest**

את המודל אימנו על סט הנתונים train default, כאשר החיפוש במרחב ההיפר-פרמטרים התבצע על האפשרויות הבאות:

min\_samples\_split = 4, 8, 16, 20

max\_depth = 5,8,10,20,30

criterion = 'gini’ or 'entropy’

'n\_estimators = 50

random\_state = 1

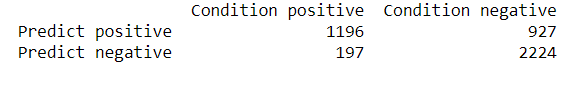
המדד שנבדק הינו AUC

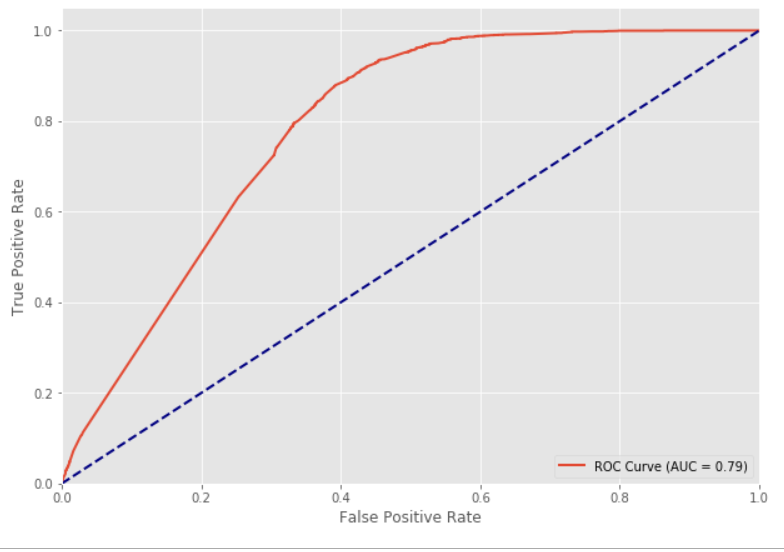
הפרמטרים שהובילו לתוצאות הטובות ביותר הינם:

כאשר מדד ה AUC שחושב הינו 0.78564

על מנת להציג גרף ROC ואת הConfusion matrix בחרנו fold רנדומאלי אימנו את המודל על סט הtrain שלו ובחנו על סט הtest שלו.

תוצאות הרצה זאת הינם:





זהו המודל בעל התוצאות הטובות ביותר עד לנקודה זאת.

**Feed Forward Neuron Network**

גם במודל זה השתמשנו במחלקה GridSearchCV אך משום שזמן האימון של רשת נוירונים עלול להיות ארוך יחסית, לא השתמשנו באופציית הגדרנו את כמות הfolds להיות 1, כלומר אימנו רק מודל אחד בכל סט פרמטרים אפשרי.

כל השכבות הינן מסוג fully contact וכוללות פונקציית אקטיבציה זהה שתיבחר בחיפוש במרחב הפרמטרים של המודל. באפשרויות לשכבות השונות השוונו בין 3 אפשרויות שונות בארכיטקטורה אך זהות במספר הפרמטרים (1000)

החיפוש במרחב ההיפר-פרמטר כלל האת האפשרויות הבאות:

Activation = ‘logistic’, ‘relu’ or ‘tanh’

Hidden layer sizes = (1000) , (500,500), (200,200,100,100,100)

Batch size = 5, 50, 200

Max\_iter = 100, 500, 2000

המדד שנבדק הינו AUC.

הפרמטרים שהובילו לתוצאות הטובות ביותר הינם:

TBD

כאשר מדד ה AUC שחושב הינו TBD

על מנת להציג גרף ROC ואת הConfusion matrix בחרנו fold רנדומאלי אימנו את המודל על סט הtrain שלו ובחנו על סט הtest שלו.

תוצאות הרצה זאת הינם TBD

# סיכום