# Ensemble Methods

*תקציר: במהלך ההכשרה, אסכם מספר נושאים "רוחביים" שאינם מתקשרים לפרק ספציפי. הנושא הרביעי הוא unbalanced dataset. במסמך זה אסביר מה זה unbalanced dataset, מה הבעיה עם זה ושיטות לפתרון הבעיה.*

## הבעיה

מידע לא מאוזן (unbalanced dataset) בו כמות המופעים מאד לא מאוזנת עבור לפחות ערך target אחד. למשל, מאגר נתונים עם 100 מופעים של מחלקה 1 ו 9900 מופעים של מחלקה 0 יהיה מאד לא מאוזן. הבעיה עם מידע לא מאוזן היא שללא טיפול מתאים, מודל ילמד כי יגיע למזעור של פונקציית הפסד (או מקסום דיוק וכו') כאשר הוא מתעלם מהמחלקה עם מעט המופעים (בדוגמה לעיל, מודל שתמיד יחזיר "0" יגיע לדיוק של 99% אבל בפועל הוא פשוט לא יודע להתמודד עם מחלקה מסוימת. כאשר יש לנו מעט דוגמאות, קשה למצוא את התפלגויות הערכים של הפיצ'רים בפועל. נרצה לפתח מודל שמצד אחד יגיע לביצועים טובים ומצד שני ידע להתמודד גם עם מופעים של מחלקות שראה מעט בזמן אימון.

## שיטות לפתרון

ישנן כמה שיטות לטיפול במידע לא מאוזן, עליהן אפרט.

### משקול ערכים של ה target

מתן משקל רב יותר למחלקה הפחות נפוצה יכול להתבטא בכמה דרכים. ראשית, ישנם מודלים שניתן לצרף להם משקול לכל מחלקה. במהלך האימון, המודל ינסה למזער יותר את הטעויות שבוצעו על המחלקה לה נתנו משקל גדול יותר. כך, ניתן לנסות לאזן בין הכמות למשקל של כל המחלקות. בנוסף, רצוי להשתמש במטריקות ממושקלות כדי שנדע מה מצב המודל בפועל (למשל, confusion matrix או weighted accuracy במקום accuracy סטנדרטי היו מראים לנו שיש בעיה במודל).

### Resampling

מאמין שזו השיטה הכי פופולרית לטיפול במידע לא מאוזן. במידה ויש לנו כמות גדולה של מידע אך הוא לא מאוזן, אפשר לנסות להשתמש בכמות קטנה יותר של מידע כך שהוא יהיה מאוזן. אם אין לנו מספיק מידע לנ"ל, אפשר לנסות לשכפל דוגמאות מהמחלקה הלא נפוצה (או לחילופין, להשתמש בשיטות ליצירת מידע מהמידע הקיים כמו SMOTE וכו'). לא נכנס לפרטים לגבי שיטות resampling כי קיים פירוט בפרק 2 בהכשרה. מודל שווה לנושא: imblearn.

### Anomaly Detection

יכול להיות ששימוש בגישה של זיהוי אנומליות הוא יותר נכון כאשר מתעסקים עם מידע לא מאוזן (לדוגמה, במקום לסווג בין פעולות משתמש לגיטימיות לכאלה שמצביעות על מתחזה ננסה לפתח מודל שיזהה אנומליות בפעולות לגיטימיות). ישנם אלגוריתמים המיועדים למציאת אנומליות (isolation forest וכו') אשר יכולים לעבוד במקרה זה.

## ניסוי

קצת מצ'וקמק, אבל השתמשתי ב digits (בערך 175 דוגמאות לכל ספרה) וחילקתי את המידע לשני סטים (אימון ובדיקה. הביצועים ההתחלתיים של המודל על סט הבדיקה מובאים באיור 1. ניתן לראות שהמודל לא מתקשה לסווג בצורה טובה אף ספרה. לאחר מכן, השארתי רק 20 דוגמאות ממחלקה 0 במידע והרצתי מחדש. כעת, ניתן לראות (איור 2) כי עבור הספרה הלא נפוצה הוא מתקשה. כעת, נאזן את המידע בעזרת oversampling (נשכפל את הדוגמאות הקיימות). באיור 3 ניתן לראות כי כעת הבעיה מוזערה והמודל מזה בצורה יותר טובה את המחלקה הלא נפוצה (מבלי לפגוע יותר מדי בביצועים על המחלקות האחרות).

|  |
| --- |
|  |

איור 1: Confusion matrix לאחר אימון על דאטה מאוזן וגם עכבר

|  |
| --- |
|  |

איור 2: Confusion matrix לאחר אימון על דאטה לא מאוזן

|  |
| --- |
|  |

איור 3:Confusion Matrix לאחר אימון על דאטה לאחר שכפול דוגמאות