# מבוא למדעי הנתונים- מטלה מס' 2

רפאל אשורוב 312054711

*\*\*\* הנתונים עליהם עבדתי הם הנתונים הנקיים מהמטלה הקודמת (אצרף קובץ CSV)*

1. תחילה אעבוד קצת על המשתנים הקטגוריאליים שלנו (Feature transformation) כדי להתאים אתם בצורה טובה יותר למודל שלנו. השתמשתי בשיטת ה-Onehot encoding ופיצלתי כל משתנה קטגוריאלי למספר משתנים בינומיאליים (0 או 1) לפי מספר הקטגוריות במשתנה.
2. עכשיו היה עלינו לבדוק את האיזון של משתנה המטרה בנתונים שלנו (death). אחרי שבדקתי את היחס בין המקרים בהם ה-death היה חיובי, חלקי כל המקרים סה"כ מצאתי שיש רק 12.4% של תוצאות חיוביות, מה שאומר שהנתונים לא מאוזנים ועלי לטפל בזה.
3. לפני שהתחלתי לבדוק את השיטות לדגימה מחדש, פיצלתי את הנתונים ל-train ו-test כדי לבחון את שיטות הדגימה בצורה יעלה יותר, ככה אבחן אותם לא רק על נתוני ה-train.
4. בחנתי את כל ארבעת השיטות כדי לסדר את חוסר האיזון: Over, Under, Over-Under (Both), Rose.  
   כדי לבחון את יכולות הניבוי של כל ארבעת הסטים של הנתונים שקיבלנו משיטות הדגימה השונות, בחרתי במודל ה-GLM הפשוט.   
   בגלל שחלק ה-test שלנו עדיין אינו מאוזן, אנחנו נצטרך ללכת עם מטריקת ה-AUC כדי לבדוק את דיוק המודל- בגלל שמטריקה זאת עובדת טוב עם חוסר איזון בנתונים.
5. לאחר בחינת התוצאות, ראינו ששיטת הדגימה Both קיבלה את ציון ה-AUC הטוב ביותר, אז בחרתי להמשיך איתה את התהליך, וביצעתי אותה על כל הנתונים לפני החלוקה.

|  |  |  |  |
| --- | --- | --- | --- |
| Over | Under | Both | ROSE |
| 0.7650497 | 0.7648216 | 0.7650976 | 0.7576163 |

1. כדי שיהיה לנו נוח יותר לעבוד עם משתנה המטרה בהמשך התהליך, אז בשלב זה, שינתי את מיקום העמודה שלו להיות העמודה האחרונה בסט הנתונים.
2. Text

   Description automatically generatedבגלל שהמשתנים הם מסדרי גודל שונים ולא מתפלגים נורמלית, אז כדי למנוע עיוות של המקדמים שלהם במודלים בצעתי scaling לנתונים.
3. אחרי ה-scaling חילקתי את כלל הנתונים:
   1. 20% מהנתונים שמרתי ל-test.
   2. את 80% מהנתונים שנשארו חילקתי ל:
      1. 70% ל-train.
      2. 30% ל-dev.

אחרי חלוקת הנתונים עשיתי בדיקה כדי לוודא ששלושת הקבוצות מאוזנות במשתנה המטרה.

1. המטריקה אותה בחרתי הייתה Accuracy. אותה בחרתי משני סיבות:

* מטריקה זו מתאימה לנתונים מאוזנים כמו שיש לי בשלב זה.
* מטריקה זו נורא פשוטה וקלה להבנה במיוחד עם משתנה מטרה בינומיאלי (כמו המשתנה שלנו), היא לוקחת את כל הרשומות בהן המודל הצליח לנבא נכון את משתנה המטרה, ומחלקת אותן בסך כל הרשומות שיש לנו (), ככה נוכל לדעת את אחוז הדיוק של המודל שלנו בצורה פשוטה ואמינה.

1. כדי לעשות את הקלסיפיקציה בחרתי ברגרסיה לוגיסטית, בגלל שמשתנה המטרה שלנו הוא קטגוריאלי בינומיאלי, ואני רוצה סה"כ לבדוק אם אדם יסווג כ-1 או כ-0.  
   כדי לבחור במודל האופטימאלי הרצתי על הנתונים את המודל הסטטיסטי של רגרסיה לוגיסטית, והמודל של Gradient Decent, וחוץ מהרצה של המודלים הרגילים הרצתי גם על כל אחד מהמודלים האלה את שלושת הרגולריזציות LASSO, Ridge, ElasticNET (בנפרד כמובן), סה"כ 8 מודלים שונים.  
   עבור המודלים עם הרגולריזציות, כדי למצוא את הלמדא האופטימלית השתמשתי ב-Cross Validation שניסה עליהם ערכים שונים של למדא עד שהגיע ללמדא האופטימלית שנותנת את הדיוק המירבי עבור המודל.
2. A screenshot of a computer

   Description automatically generated with medium confidenceלפנינו תוצאות ה-Accuracy של המודלים השונים. אפשר לראות שיחס הניבוי די דומה בין כל המודלים, בינתיים נראה כי כנראה לא תהיה לנו בעיה של Over\Under fitting בגלל שהתוצאות דיי דומות בין כל המודלים.  
   המודל שקיבל את הציון בגבוהה ביותר הוא המודל של הרגרסיה הלוגיסטית עם המטריקה של Ridge, נמשיך עם המודל הזה ל-test.
3. תוצאות המודל על נתוני ה-test:



קודם כל ניתן לראות שגם ב-test קיבלתי יחס ניבוי כמו ב-dev וב-train, מה שאומר שאין לנו בעיה של Over\Under fitting.  
חוץ מזה הגענו לרמת ניבוי של אזור ה-70%, מה שאומר שהמודל שהגענו אליו הוא לא רע, אבל הוא ממש לא מושלם. אני יכול לחשוב על כמה סיבות למה לא הגענו לאחוזי דיוק גבוהים יותר:

* הנתונים שעבדנו איתם לא היו מאוזנים, אז אולי אם הייתי משתמש בשיטת דגימה אחרת (לא Both) הייתי יכול לקבל אחוזי דיוק יותר גבוהים.
* אולי טעיתי בחלק מהשלבים במתלה הקודמת: ב-EDA, או בניקוי הנתונים, או בבחירת המשתנים, או בטיפול בערכי הקיצון, או בטיפול בערכים החסרים... כל טעות אי שם בדרך הייתה יכולה לשבש את אחוזי ההצלחה של המודל שלנו. אולי השארתי רעש בנתונים שהרס את הדיוק, או אולי הפוך בחרתי במשתנים לא נכונים, מה שהיה יכול ליצור רעש בנתונים. אולי בכלל בבחירת המשתנים השמטתי משתנים היה להם ערך בדיוק הניבוי.
* סיבה אפשרית נוספת, אולי מראש הנתונים לא היו נכונים והיה חוסר דיוק באיסוף שלהם. דבר זה יכול לגרום לרעש ופגיע בדיוק.