



סקירה של כשלים בבנייה וביישום מודלים של חיזוי אנליטי



יעקב זהבי

פרופ' יעקב זהבי הוא פרופסור אמריטוס בפקולטה לניהול ע"ש קולר באוניברסיטת תל אביב. הוא אחד מפורצי הדרך בתחום כריית המידע (Data Mining) בעולם נתוני העתק, שבו הוא מעורב במספר חזיתות — מחקר, הוראה, פיתוח תוכנה ויישומים לקבלת החלטות. פרופ' זהבי החל את הקריירה המקצועית שלו בתחום של מערכות מידע בתור מנתח מערכות בסקטור הציבורי. עם סיום לימודי הדוקטורט באוניברסיטת Pennsylvania הצטרף לפקולטה לניהול באוניברסיטת תל אביב ובמשך מספר שנים עסק בפיתוח וביישום של מודלים של חקר ביצועים וקבלת החלטות בתחום האנרגיה והחשמל. בסוף שנות השמונים עבר "הסבה מקצועית" לתחום של שיווק מבסיסי נתונים, וממנו הגיע לתחום של כריית מידע שבו הוא עוסק עד היום. זכה פעמיים רצופות במדליית הזהב בתחרות השנתית לגילוי ידע (Knowledge Discovery) שמאורגנת על ידי (American ACM Computation Machinery). מספר מאמרים שלו בתחום זה זכו בפרסים על מצוינות אקדמית.

תקציר

במאמר זה אנו סוקרים מספר כשלים בבנייה וביישום של מודלים לחיזוי אנליטי, בדגש על מודלים של חיזוי מבוססי רגרסיה בתחום של שיווק ישיר. חילקנו את הדין בכשלים אלה לשלוש קטגוריות: כשלים הנובעים מהטיות בבניית מודל החיזוי, כשלים הנובעים מהכנת הנתונים למודל החיזוי, וכשלים הנובעים מיישום מודל החיזוי לקבלת החלטות. כשלים אלה מתבטאים בשונות גדולה בתוצאות העסקיות של מודל החיזוי שאותה הדגמנו באמצעות הצגה של אירוע אמיתי (תחרות ה-1998 KDD CUP), שממנו ניתן ללמוד על ההבדלים הגדולים בתוצאות העסקיות של מודלים שונים לחיזוי אנליטי, על אף שהם התבססו על אותם בסיסי נתונים. המסקנה העיקרית של הדין במאמר היא שבניית מודל חיזוי בעולם נתוני העתק אינה בעיה "אוטומטית" של הפעלת תוכנת חיזוי, אלא מחייבת להתמודד עם הכשלים השונים בבניית מודלים לחיזוי. עם זאת, ההשקעה הדרושה על מנת להתמודד עם הכשלים בבניית מודל חיזוי איכותי, יציב ומשמעותי היא מאוד משתלמת, שכן היא מניבה תשואה עסקית גבוהה יותר שמתבטאת לא רק בתמורה כספית גבוהה אלא גם בעוד הרבה יתרונות איכותיים.

1. הקדמה

והיישום שלהם בתחום של שיווק מבסיסי נתונים (Database marketing) עם דגש על שיווק ישיר (Direct marketing). ביישומים של שיווק ישיר המטרה היא להציע מוצרים ושירותים באמצעות פנייה ישירה ללקוחות בבסיסי הנתונים של הארגון. אבל במקום לפנות לכל מיליוני הלקוחות בבסיס הנתונים, בידיעה שרק חלק קטן מאוד מהם מתכוון להיענות להצעה השיווקית, המטרה היא לחזות את שיעורי התגובה של לקוחות חדשים על בסיס שיעורי תגובה ידועים של לקוחות שנחשפו להצעה לרכוש את המוצר/שירות בעבר, ולפנות רק ללקוחות ששיעור התגובה החזוי שלהם עובר סף מסוים שנקבע על סמך שיקולים כלכליים או אחרים. מתוך מגוון ערוצי השיווק הישיר, אנו נתמקד במאמר זה בשיווק באמצעות הדואר, ולשם פשטות נניח שההצעה השיווקית מתייחסת למוצר (או שירות) יחיד (Solo mailing). היתרון של מבצעי שיווק ישיר המבוססים על חיזוי אנליטי מתבטא בעיקר בחיסכון גדול בעלויות השיווק, שכן היא מאפשרת לנפות ממבצע השיווק את כל מאות אלפי הלקוחות שעל פי מודל התגובה לא מתכוונים להיענות להצעה השיווקית.

התחום של חיזוי אנליטי הוא אחד התחומים ה"חמים" היום בכריית מידע. בעיות חיזוי הן בעיות שבהן קיים קשר תוצאתי בין המשתנה התלוי למשתנים המסבירים, שמאפשר לנו לחזות את משתנה הפלט כפונקציה של משתני הקלט. תחום החיזוי האנליטי מכיל מגוון רחב של מודלים ושיטות לחיזוי, ומביניהם נתמקד במאמר זה במודלים לחיזוי מבוססי רגרסיה, שהם ללא ספק המודלים הנפוצים והוותיקים ביותר לחיזוי. האירוע שאנו רוצים לחזות הוא משתנה התגובה (Response variable), או המשתנה התלוי, שאנו מנסים להסביר באמצעות סדרה של משתנים מסבירים המכונים גם משתנים בלתי תלויים (Independent variables), פרדיקטורים (Predictors) או מאפיינים (Attributes, features).¹ למשל, בתחום של שיווק ישיר המשתנים המסבירים משקפים קניות שהלקוח עשה בעבר, פניות שנעשו ללקוח לרכוש מוצרים ושירותים שלא נענו, היסטוריית תשלומים ואשראי של הלקוח, מאפיינים דמוגרפיים כגון גיל, מצב משפחתי, אשכול סוציו-אקונומי ואחרים.

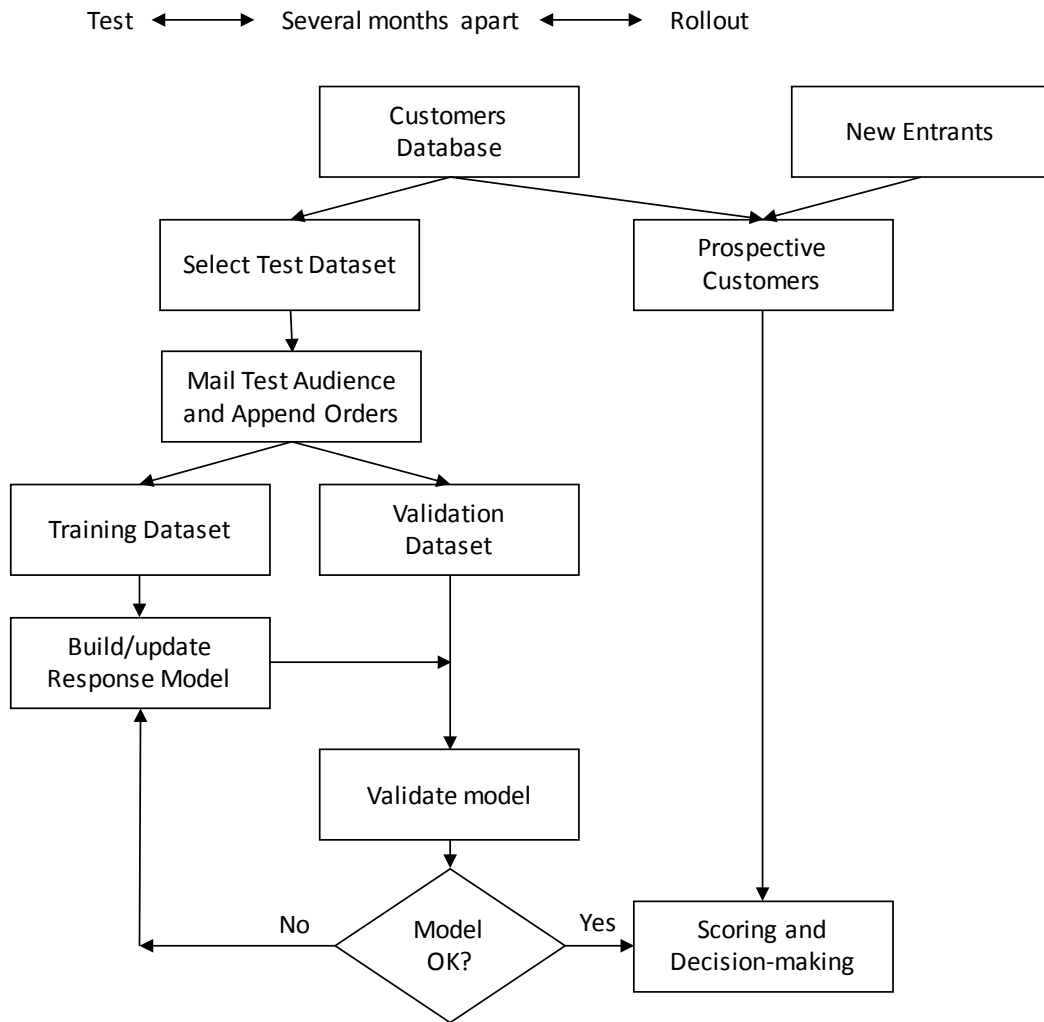
אנו נחלק את הדיון בכשלים בבניית מודלים לחיזוי אנליטי לשלוש קטגוריות: כשלים הנובעים מהטיות בבניית המודל, כשלים שמקורם מהטיות בנתונים, וכשלים הנובעים מהטיות ביישום המודל לצורך קבלת החלטות. השונות בתוצאות

הכמות העצומה של נתונים המגיעים כיום מכל עבר – רשתות חברתיות, בלוגים ופורומים באינטרנט, IoT, חברות עסקיות, רשויות ממשלתיות, עיתונים מקוונים ולא מקוונים, חיישנים ומצלמות מסוגים שונים (וורק רשימה חלקית), הביאה לפיתוח מואץ של מודלים אנליטיים כדי לסנן את המוץ מהתבן ולהפיק מידע איכותי מהנתונים לצורך קבלת החלטות. התחום העוסק בפיתוח של מודלים אנליטיים בעולם נתוני העתק הוא התחום של כריית מידע (Data mining), שלצורך העניין כולל בתוכו גם נישות של למידת מכונה (Machine learning) ובינה מלאכותית (Artificial intelligence). ואומנם, הספרות המקצועית הדנה בתחום של כריית המידע היא רחבה וענפה וכוללת מגוון של נישות ומודלים להתמודדות עם בעיות עסקיות שונות. עיקר העיסוק המחקרי בתחום הוא לפתח מודלים עם יכולת עיבוד מהירה שממזערים טעויות ומניבים תוצאות מדויקות יותר על מנת לשפר את תהליך קבלת ההחלטות. לאחרונה מושם דגש רב על בניית מודלים בצורה אוטומטית על מנת להנגיש את הטכנולוגיה הזו לכלל מקבלי ההחלטות גם בארגונים עסקיים בינוניים וקטנים שאין להם את היכולת והמשאבים להתמודד עם בניית מודלים אנליטיים מתוחכמים ורב־ממדיים.

אבל האיכות והדיוק של המודלים האנליטיים תלויים במידה רבה לא רק בטיב המודל ובמידת התאמתו לבעיה העסקית, אלא גם באיכות ובטיב הנתונים המוזנים למודל, כמאמר הפתגם "Your model is only as good as your data". אך משום מה הספרות המקצועית בתחום דנה יותר בבנייה ופתרון של מודלים אנליטיים ורק לאחרונה החלו להופיע פרסומים העוסקים גם בקשרי הגומלין בין טיב הנתונים לאיכות המודל, ובהם (Ying (2019), Kim et al. (2019). על אף שבמאמר זה אנו עוסקים בעיקר בכשלים הנוגעים למודלים מבוססי רגרסיה, ראוי לציין גם פרסומים אחרים המתייחסים לתחומים חדשים ומעניינים, כגון פתרון הכשלים הללו במודלים מבוססי רשתות נוירונים (Li et al. (2019), Lim et al. (2020), ואחרים.

במאמר זה נסקור מספר כשלים נפוצים בבניית מודלים הנובעים מסוגיות הנוגעות גם לבניית המודל וגם להכנת הנתונים למודל. מודלים שונים, ולפעמים גם יישומים שונים, דורשים התייחסות שונה לנתונים. לכן במאמר הזה נתייחס בעיקר למודלים של חיזוי אנליטי (Predictive analytics)

¹ במאמר זה נתייחס למושגים מאפיינים, משתנים מסבירים, פרדיקטורים ומשתנים בלתי תלויים כמושגים נרדפים.

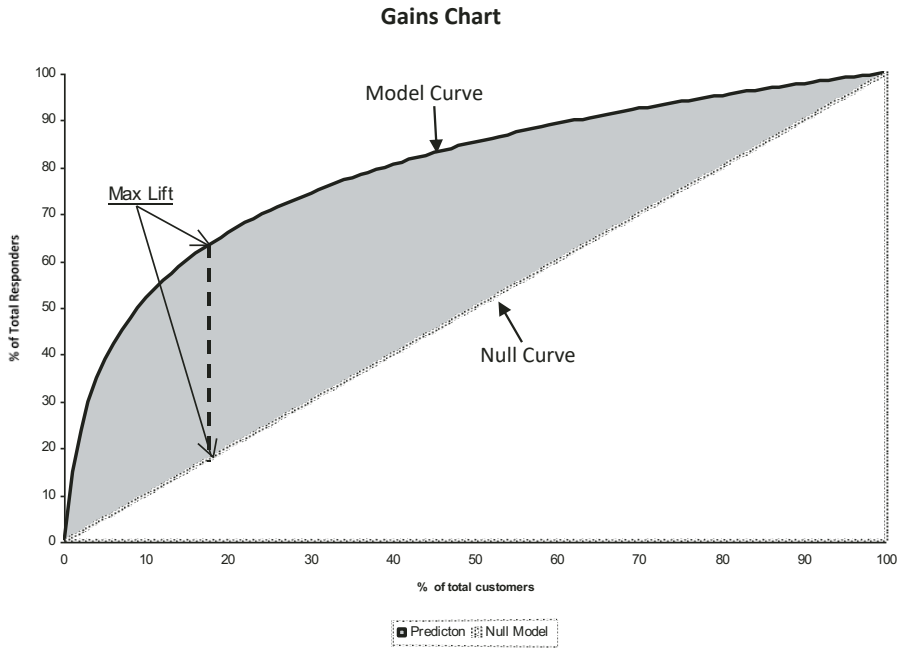


2. תיקוף ובדיקה של מודלים לחיזוי אנליטי

תהליך החיזוי האנליטי מורכב למעשה משלושה שלבים עיקריים:

- בניית מודל תגובה (*Response modeling*) המתאים מודל לנתונים.
- תיקוף ובדיקה של המודל.
- תהליך של מתן ציון (*Score*) לכל לקוח המשמש לקבלת החלטות.

העסקיות של מודלים שונים של חיזוי אנליטי על אותו קובץ נתונים, שככל הנראה נובעת מהכשלים הנדונים במאמר הזה, יכולה להיות גדולה מאוד ונעה בין רווחים לארגון ובין הפסדים לארגון. אנחנו נדגים זאת באמצעות אירוע אמיתי המבוסס על התחרות *KDD CUP 1998*. המטרה של המאמר הזה היא לא בהכרח להציע פתרונות לכשלים אלה, אלא להציף את הבעיות שנובעות מהן על מנת ליצור מודעות בקרב המשתמשים העסקיים ומפתחי המודלים לכשלים האלה ולהסב את תשומת ליבם לצורך למצוא פתרונות מתאימים על מנת לשפר את תהליך קבלת ההחלטות.



מאחר שאנו מתמקדים במאמר הזה בכשלים בבנייה וביישום של מודלים לחיזוי אנליטי המשפיעים על טיב המודל ואיכות החיזוי, יש חשיבות רבה לתקף את המודל ולבדוק שהוא בעל יכולת הכללה (Generalizable) ושאפשר להפעיל אותו גם על הלקוחות החדשים. בסעיף זה נדון בקצרה במדדים להערכת איכות המודל, בדגש על מודלים של רגרסיה. דיון נרחב יותר בנושא מופיע אצל זהבי (2017).

הספרות המקצועית מציעה מספר מדדים להעריך את איכות המודל — מדדים "גלובליים" שמבטאים את טיב ההתאמה של המודל לנתונים, ומדדי רווחיות המבטאים את יכולתו של המודל להפריד בין הלקוחות הרווחיים והבלתי רווחיים. בקבוצת המדדים הגלובליים ניתן למנות את מקדם הקביעה (*Coefficient of determination*), הידוע גם בתור R^2 (*R-square*), מדדים המבוססים על סכום הסטיות הריבועיות בין הערכים האמיתיים של המשתנה התלוי והערכים החזויים שלו על פי המודל, כגון *RMSE - Root Mean Square Errors*, ועוד. מדדי הרווחיות מיינים את הלקוחות בסדר יורד של תחזיות המודל בטבלת רווחים (*Gains table*) ברמה של אחוזונים, לרוב עשירונים, מהעשירון העליון לעשירון התחתון, ומונים את מספר המגיבים בכל עשירון. מודל "טוב" הוא מודל

תרשים הזרימה של תהליך החיזוי האנליטי המתבסס על דיוור מבחן מופיע באיור 1.

בבעיות שיווק, מודל התגובה מבוסס על סמך מדגם של לקוחות מתוך בסיס הנתונים של הארגון שנחשפו למוצר המוצע בעבר, חלקם הקטן הניבו ורכשו את המוצר וחלקם הגדול לא הניבו ולא רכשו את המוצר. תרשים הזרימה באיור 1 מתייחס למקרה של מוצר (או שירות) חדש שאין עליו כל מידע על תגובת הצרכנים. במקרה הזה מודל התגובה מתבסס על דיוור מבחן (*Test mailing*) שנבחר באופן מקרי מתוך בסיס הנתונים של הלקוחות. לאחר בנייה ותיקוף המודל, מפעילים אותו על הקובץ שמכיל את יתרת הלקוחות בבסיס הנתונים שלא לקחו חלק בבניית מודל התגובה, בתוספת לקוחות שהצטרפו למערכת בטווח הזמן שבין דיוור המבחן למבצע השיווק המלא (אלה הלקוחות הפוטנציאליים — *Prospective customers* — שאותם נכנה מכאן ואילך בשם "לקוחות חדשים"). על מנת לתת ציון לכל לקוח שמשקף את שיעורי התגובה שלהם. תלוי בסוג המודל, הציון הזה מייצג את הסתברות התגובה של הלקוחות, או את מדרג התגובה שלהם בסולם 0-1, או 1-10. ציונים אלה משמשים לצורך קבלת החלטות.

שמצליח "לתפוס" את מרבית המגיבים בעשירונים העליונים ואת מיעוטם בעשירונים התחתונים.

מאחר שמודל רגרסיה הוא למעשה מודל אופטימיזציה שמטרתו למצוא את מקדמי הרגרסיה שמשיאים פונקציית מטרה מסוימת, לדוגמה מקדם הקביעה המותאם (*Adjusted R-square*) ברגרסיה ליניארית, זה לא מפתיע שהמודל נותן תוצאות טובות על קובץ הנתונים ששימש לצורך בניית המודל. לכן כדי לבדוק את טיב המודל ואיכות החיזוי שלו צריך לבדוק את המודל על קובץ נתונים שלא לקח חלק בבניית המודל. הדרך המומלצת היא לפצל את קובץ נתוני המבחן (*Test dataset*) באופן אקראי לשני קבצים בלתי תלויים — קובץ אימון (*Training dataset*) וקובץ תיקוף (*Validation dataset*), כמתואר באיור 1, לבנות את מודל התגובה על בסיס קובץ האימון ואז להפעיל אותו על קובץ התיקוף כדי לחזות את הערך של המשתנה התלוי עבור כל תצפית. מאחר שקובץ התיקוף מכיל גם את התגובות בפועל, ניתן להשוות את התחזיות של מספר המגיבים של המודל עם מספר המגיבים בפועל ברמה של העשירונים בין קובץ האימון לקובץ התיקוף. אם התפלגות מספר המגיבים בפועל ברמה של העשירונים בין שני הקבצים הוא דומה, הדבר מעיד על מודל יציב. מבחינה מעשית אפשר להשתמש ב"כלל אצבע" שלפיו הפרשים של 5%–10% בין תחזית מספר המגיבים בקובץ התיקוף ובין מספר המגיבים בפועל בקובץ האימון הם עדיין הפרשים "סבירים", המעידים על מודל יציב ובעל יכולת הכללה.

תמונה חזותית של איכות מודל החיזוי ניתן לקבל באמצעות תרשים הרווחים (*Gains chart*) באיור 2, שמציע למעשה הצגה גרפית של טבלת הרווחים. הציר האופקי (ציר X) בתרשים הרווחים מציין את האחוז המצטבר של הלקוחות, והציר האנכי (ציר Y) את האחוז המצטבר של המגיבים, כאשר התצפיות ממוינות בסדר יורד של תחזית התגובה. תרשים הרווחים מציג את התוספת לרווחים שמתקבלת באמצעות מודל החיזוי, המיוצג באמצעות עקומת המודל (*Model curve*) בחלק העליון של התרשים, לעומת מודל ה"אפס" (*Null model*) שמניח שכל הלקוחות שווים, המיוצג באמצעות הקו הישר בזווית של 45 מעלות היוצא מהראשית (*null curve*). על פי תרשים הרווחים באיור 2, אם היינו פונים ל-20% מהצרכנים על פי מודל האפס (כלומר באופן מקרי), היינו מצפים "לתפוס" בממוצע 20% מהמגיבים. לעומת זאת, אם פונים ל-20% מהלקוחות הטובים ביותר על פי מודל החיזוי, אנו מצפים "לתפוס" כ-60% מהמגיבים, שיפור משמעותי לעומת מודל

האפס. ככל שהמרחק בין עקומת המודל ובין מודל האפס גדול יותר, כך מודל החיזוי הוא "טוב" יותר, כלומר מצליח להבחין יותר טוב בין המגיבים ללא מגיבים.

מדד נוסף לאיכות המודל הוא סטטיסטי המקסימום, הידוע גם בשם *Kolmogorov Smirnov*, ומבטא את המרחק המקסימלי (*Max-lift*) בין עקומת המודל לבין הקו הישר המיצג את מודל האפס. ככל שהערך של סטטיסטי המקסימום גדול יותר, כך איכות החיזוי של המודל טובה יותר. ישנם גם מבחנים סטטיסטיים מדויקים יותר המתבססים על סטטיסטי המקסימום כדי לבדוק אם יש הבדל משמעותי (*Significant*) בין עקום המודל לעקום האפס (*DeGroot, 1993*). לסטטיסטי המקסימום יש גם יתרון נוסף, שכן הוא מאפשר להשוות בקלות, באמצעות מדד אחד, את איכות החיזוי של מודלים שונים.

3. כשלים הנובעים ממודלים מוטעים (misspecified models)

3.1 הטיות בבניית מודלים לחיזוי מבוססי רגרסיה

במודל תגובה המבוסס על רגרסיה, המטרה היא ל"הסביר" את המשתנה התלוי Y באמצעות פונקציה של J משתנים מסבירים. X_1, X_2, \dots, X_j

$$Y = f(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_j X_j) + \varepsilon$$

כאשר:

$\beta_0, \beta_1, \beta_2, \dots, \beta_j$ המקדמים של המשתנים המסבירים (הפרמטרים).

$f(\cdot)$ פונקציה כלשהי של המשתנים המסבירים.

ε – הפרעה מקרית (שניאה).

בעיות חיזוי יכולות להיות משני סוגים — אמידה (*Estimation*) או סיווג (*Classification*). בבעיות אמידה המשתנה התלוי הוא רציף, למשל כמות הכסף שלקוח יוציא על קניות מקטלוג, נובה התרומה של לקוח לארגון צדקה, גודל התביעה בגין תאונת רכב, ועוד, והמטרה היא לאמוד את התוחלת של המשתנה הזה עבור כל לקוח. במודל זה המשתנה התלוי Y הוא משתנה רציף ומודל הרגרסיה הוא פונקציה ליניארית.

בבעיות סיווג המשתנה התלוי הוא בדיד, שלרוב הוא בינארי עם הערכים 0/1 (כן/לא). לדוגמה, 1 – אם הלקוח הזמין את המוצר, 0 – אם לא הזמין, והמטרה היא לחזות את הסתברות הרכישה של המוצר עבור כל לקוח. במודל זה המשתנה התלוי Y הוא בינארי 0/1 ומודל הרגרסיה הוא פונקציה לוגיסטית.

את מקדמי הרגרסיה מוצאים באמצעות פתרון בעיית אופטימיזציה. במאמר זה נדלג על השיטות לפתרון בעיות האופטימיזציה, אבל אלה נדונות בהרחבה בספרות, למשל בספרם הפופולרי של (Hastie et al., 2009), שגם הראו שניתן לאמוד את הפרמטרים של מודל רגרסיה לוגיסטית באמצעות פתרון של סדרה של בעיות רגרסיה לינארית עם משקלות. יוצא מכך שמודל הרגרסיה הליניארי מהווה את הבסיס התיאורטי גם למודל הרגרסיה הלוגיסטית.

הבעיה המורכבת ביותר בבנייה של מודל תגובה היא בחירת המשתנים המשפיעים ביותר למודל, בדרך כלל כמה עשרות בודדות של משתנים מתוך האוסף הגדול מאוד של משתנים מסבירים פוטנציאליים. אנשי כריית המידע מכנים בעיה זו כבעיה של בחירת מאפיינים (Feature selection), ואילו הסטטיסטיקאים כבעיית ספציפיקציה (specification). בחירה נכונה של המשתנים המשפיעים במודל חיזוי אנליטי רב-ממדי קובעת במידה רבה את האיכות והדיוק של מודל החיזוי. הגישה המובילה לבחירת המשתנים במודל רגרסיה מתבססת על תהליך סטטיסטי של בדיקת השערות (Test of hypothesis), שלפיה מכניסים למודל משתנים מסבירים שעוברים את סף רמת המובהקות (Level of significance) שנקבעת מראש, בדרך כלל 5% או 1%.

אומדן מדויק של הפרמטרים ניתן לקבל רק אם בונים את המודל על סמך כלל אוכלוסיית היעד. אולם בשל הממדים הגדולים של בסיסי הנתונים, הן מבחינת גודל קובצי הנתונים שמכילים לעיתים מיליונים של לקוחות, והן מבחינת המשתנים המסבירים במודל שמספרם עשוי להגיע לכמה אלפים ויותר, מקובל לבנות את מודל התגובה על סמך מדגם מייצג של האוכלוסייה. היתרון של מדגם מייצג הוא בקיצור משמעותי של תהליך בניית המודל, אבל מצד שני הוא משפיע על הדיוק והאיכות של המודל בשל טעויות הדגימה: מחד, טעות מסוג ראשון (Type-I error) שמשמעותה הכנסת משתנים בלתי מובהקים למודל, ומאידך טעות מסוג שני (Type-II error) שמביאה להשמטה של משתנים מובהקים מהמודל.

לטעויות אלה בבניית מודל החיזוי יש משמעות עסקית רבה. למשל, בהקשר של בעיות בשיווק ישיר שמטרתן לאתר לקוחות שישתתפו במבצע השיווק ("טרנט"), טעות מסוג ראשון עשויה למנוע מלקוחות רווחיים להשתתף במבצע השיווק, ואילו הטעות מסוג שני עשויה לשתף לקוחות בלתי רווחיים במבצע השיווק. שני סוגי הטעויות כרוכים בהפסדים לחברת השיווק – הטעות מסוג ראשון גורמת להפסד רווחים בנין השמטה של לקוחות רווחיים במבצע השיווק וכן בהפסד מוניטין, ואילו הטעות מסוג שני כרוכה בהפסד כספי "ממשי" בנין העובדה של לקוחות בלתי רווחיים לוקחים חלק במבצע השיווק.

למרבה הצער, אי אפשר למנוע את הטעויות מסוג ראשון ושני אלא אם בונים את המודל על סמך כל האוכלוסייה, דבר שהוא לא מעשי. אבל הספרות מציעה מספר שיטות להקטין את ההסתברות לטעויות, ובהן שימוש במדגמים גדולים יותר, שליטה על הטעויות מסוג ראשון ושני באמצעות תיקון בונפרוני (Bonferroni Adjustment), שיטת התגליות השנויות (FDR – False Discovery Rates) (Benjamini and Hochberg, 1995), וכן שימוש בנישוחות שונות ש"קונסות" את פונקציית המטרה על סמך מספר המשתנים המסבירים שכבר נמצאים במודל, בשאיפה להקטין את מספר המשתנים המסבירים שנכנסים למודל, כגון $AIC - Akaike Information Criterion$ (Akaike, 1973), $BIC - Bayesian Information Criterion$ (Schwarz, 1978), ואחרות.

אבל גם אם עוברים על כל המשתנים המסבירים אחד אחד, ומחליטים על סמך מבחן סטטיסטי האם להכניס או להשמיט את המשתנה ה"תורן" מהמודל, עדיין לא נקבל בהכרח את המודל המיטבי, שכן טיב המודל נקבע לא רק על סמך התהליך הסטטיסטי אלא גם על סמך הסדר של המשתנים המסבירים שמשותפים בתהליך ומספרם. כמו כן, קיימת האפשרות שלאחר שמכניסים משתנה מובהק למודל, אנו עשויים להוציא אותו מהמודל בשלב מאוחר יותר בנין הכנסת משתנה אחר שהפך את המשתנה הזה לבלתי מובהק. לכן על מנת לקבל את המודל המיטבי ביותר, יש צורך לכאורה לחזור על תהליך בדיקת ההשערות על כל הקומבינציות האפשריות של המשתנים המסבירים, משימה שהיא כמובן בלתי אפשרית. מגוון של פתרונות יריסטיים פותחו על מנת להתמודד עם בעיית בחירת המשתנים המסבירים במודל רגרסיה רב-ממדי, והנפוצה שבהם היא גישת הרגרסיה בצעדים (SWR – stepwise regression) (Efoymson, 1960). פרט לשיטת הרגרסיה בצעדים, קיימות

עוד מגוון שיטות לבחירת משתנים מסבירים במודלים של חיזוי שנדונות בהרחבה בספרות (למשל, Miller, 2002).

בחירה לא נכונה של משתנים מסבירים למודל התגובה, בשילוב עם הטעויות מסוג ראשון ושני, הנה ללא ספק המקור העיקרי להטיות בבניית מודל התגובה. זיהוי מודלים מוטים, תוך שימוש במדדים שפורטו לעיל ובמדדים אחרים הנדונים בספרות, מהווה חלק משמעותי בתהליך נילוי הידע מכיוון ששימוש במודל מוטה לקבלת החלטות עשוי לעלות לארגון בהפסדים ניכרים — כספיים ואחרים.

3.2 התאמת יתר (Over fitting)

המטרה העיקרית של בניית מודל תגובה בבעיית שיווק היא לסנן את הלקוחות הלא רווחיים ממבצע השיווק. מצב של התאמת יתר נוצר כאשר המודל נותן תוצאות טובות על קובץ הנתונים ששימש לבניית המודל, אך נותן תוצאות גרועות כאשר מיישמים את המודל על לקוחות חדשים. התאמת יתר היא "מכה" אופיינית במודלים של חיזוי רב-ממדיים. הבדלים משמעותיים בין עקומות המודל של קובץ האימון וקובץ התיקוף בתרשים הרווחים של איור 2 יכולה להעיד על התאמת יתר. ברור שלא ניתן להשתמש במודל עם התאמת יתר לצורך קבלת החלטות.

התאמת יתר היא לרוב תוצאה של מודלים מוטים, בין אם בגין הכנסת משתנים בלתי מובהקים למודל ובין אם בשל השמטה של משתנים מובהקים מהמודל. סיבות אחרות הן מיעוט אינפורמציה שלא מאפשרת לאמוד מודל "טוב", כגון קובץ אימון שמכיל מספר קטן מאוד של מגיבים יחסית למספר הלא מגיבים, מספר רב של משתנים מסבירים במודל אבל יחסית מספר מועט של תצפיות, משתנים מסבירים שגויים, רעשים מקריים, ובעיות דאטה אחרות (שחלקן יידונו להלן).

אין "מרשם" מסודר כדי לפתור בעיות של התאמת יתר. הכלים המרכזיים הם לחזק את המודל באמצעות הגדלת מדגם האימון, הוספת משתנים מסבירים למודל ו/או טרנספורמציות שלהם, בניית מודל תגובה "מדולל" (Parsimonious) יותר, העשרה של נתוני הלקוחות ממקורות נוספים, חלקם אפילו מחוץ לארגון, כגון נתונים דמוגרפיים, נתוני אשראי, נתוני סגנון חיים ועוד. Ying (2019) דן בחלק מהשיטות האלה בהרחבה. כמו כן ניתן לנסות גם מודלים אחרים של חיזוי

אנליטי כדי לפתור את הבעיה, כגון רשתות עצביות (Neural networks), עצי החלטה (Decision trees) ואחרים. גישה נוספת שזוכה לפופולריות רבה לאחרונה היא גישת תכול (Ensemble) שמשלבת אומדים של המשתנה התלוי (Scores) ממספר מודלים על מנת להקטין את שונות (Variability) התחזית. הגישות המובילות הן Bagging (Brieman, 1996) ו-Boosting (Friedman et al., 1998).

3.3 התאמת חסר (Under fitting)

התאמת חסר היא "תמונת ראי" של התאמת יתר. התאמת חסר מתייחסת למקרה שבו מודל התגובה אינו מסוגל להבחין בבעיית שיווק בין קונים פוטנציאליים ללא קונים. אפשר לזהות התאמת חסר באמצעות תוצאות המודל על קובץ האימון. שיעורי תגובה לא מונוטוניים כפונקציה של העשירונים במדגם האימון, עקומת מודל במדגם האימון שקרובה מדי למודל האפס, כמו גם הבדלים קטנים מאוד בשיעורי התגובה בין העשירון העליון והתחתון, יכולים להעיד על התאמת חסר. ישנן מספר סיבות להתאמת חסר: מודלים מוטים, טרנספורמציות שגויות של משתנים מסבירים, השמטה של משתנים מסבירים "חשובים", וסיבות אחרות. גם כאן אין מרשם ברור כדי לפתור את בעיית התאמת החסר. כמה אפשרויות הן: לנסות מודלים אחרים של חיזוי, חלוקה של אוכלוסיית היעד למספר סגמנטים "הומוגניים" ובניית מודל תגובה נפרד לכל סגמנט, העשרה של סט הנתונים ששימש לבניית מודל התגובה ממקורות פנימיים וחיצוניים (כגון נתונים דמוגרפיים, משתני סגנון חיים, נתוני אשראי ועוד), שימוש במדגמים גדולים יותר לצורך בניית מודל התגובה, ועוד. ברוב המקרים של התאמת חסר מדובר בתהליך שעשוי להיות מייגע ודורש מיומנות ויצירתיות.

3.4 מדגמים מבוססי בחירה (Choice based samples)

כאמור, מודל התגובה בבעיות שיווק מתבסס על מדגם מקרי מאוכלוסייה שנחשפה למוצר בעבר. אולם מכור שיעורי התגובה בבעיות שיווק ישיר הם קטנים מאוד, לעיתים פחות מחצי אחוז (ובמצעי שיווק אינטרנטיים אף הרבה פחות מזה), מה שמביא לכך שמספר המגיבים במדגם המשמש לבניית מודל התגובה כולל לרוב מספר קטן מאוד של מגיבים יחסית למספר הלא מגיבים, מה שלא מאפשר בניית מודל תגובה "טוב". הפתרון

אנחנו נמשיך לדון להלן בהטיות אלה מנקודת הראות של מודלים לחיזוי בעולם העסקי, בדגש על עולם השיווק.

4.1 יחסים לא ליניאריים (Non-linear relationships)

מודלים של רגרסיה מניחים לרוב שיש קשר ליניארי בין המשתנה התלוי ובין כל אחד מהמשתנים המסבירים, אבל זה לא בהכרח המצב במציאות. להיפך, ברוב המקרים הקשר בין המשתנים הוא לא ליניארי. ניקח לדוגמה את הקשר בין רמת ההכנסה (משתנה תלוי) לבין גיל הלקוח (משתנה מסביר). לרוב רמת ההכנסה עולה כפונקציה של הגיל עד שהיא מגיעה לרמה מרבית, ואז היא הולכת ויורדת עם הגיל. כלומר מדובר כאן על קשר לא ליניארי, קרוב לוודאי קשר פרבולי. התעלמות מהעובדה שהקשר בין משתנים הוא לא בהכרח ליניארי הוא אחד הגורמים להטיה במודלים לחיזוי. הדרך המקובלת לבטא יחסים לא ליניאריים במודל רגרסיה היא באמצעות טרנספורמציות של משתנים. למשל, בדוגמה של הקשר הפרבולי לעיל, טרנספורמציה רלוונטית היא פונקציה פולינומית מהסוג $y = x^\alpha$, כאשר $-2 \leq \alpha \leq 2$. אם פרמטר של הפונקציה הפרבולית שמוגדר מראש. אם $\alpha < 1$, הטרנספורמציה ממתנת את ההשפעה של המשתנה המסביר X (משתנה הגיל בדוגמה הקודמת) על המשתנה התלוי Y (רמת ההכנסה); וההיפך, אם $\alpha > 1$, הטרנספורמציה מעצימה את ההשפעה של המשתנה המסביר X על המשתנה התלוי Y . עבור המקרה הספציפי $\alpha = 0$, הטרנספורמציה מוגדרת בתור $Y = \log(X)$.

היתרון של הטרנספורמציה הפרבולית הוא במגוון הרחב של האפשרויות להגדיר קשרים לא ליניאריים בין משתני הקלט והפלט. החיסרון של הטרנספורמציה הזו הוא הצורך להגדיר מראש את צורת הפונקציה. ומה אם הקשר האמיתי בין המשתנה התלוי למשתנה המסביר הוא דמוי פרבולי ולא תואם בדיוק את הפונקציה הפרבולית? גישות אלטרנטיביות להגדרת יחסים לא ליניאריים בין משתנה פלט למשתנה הקלט הן גישות המבוססות על הנתונים עצמם. מועמדים אפשריים הם פונקציית מדרגה (Binning, step function) ופונקציה ליניארית למקוטעין (Piece wise linear function). במקרה של פונקציית מדרגה מחלקים את תחום ההשתנות של המשתנה המסביר X למקטעים זרים וממצים (mutually Exclusive and exhaustive intervals), למשל על פי רבעונים

לבעיה הזו היא לכלול במדגם האימון פרופורציה גדולה יותר של מניבים מאשר הפרופורציה שלהם באוכלוסייה, לעיתים אף את כל המניבים באוכלוסייה ורק מדגם קטן של לא מניבים. למעשה מדובר כאן על מדגם שכבות (Stratified sampling) עם שתי שכבות — אחת של מניבים ואחת של לא מניבים. בהקשר של בעיות בשיווק, מדגמים אלה נקראים מדגמים מבוססי בחירה שכן הם מבוססים על הבחירה (choice) של הלקוח אם להיענות או לסרב להצעה לרכוש את המוצר המוצע (Ben Akiva and Lerman, 1987). עם זאת, מודל תנובה שמתבסס על מדגמים מבוססי בחירה אינו משקף את כלל אוכלוסיית היעד אלא את אוכלוסיית המדגם. למשל, מודל רגרסיה לוגיסטית המבוסס על מדגם הכולל פרופורציה גדולה יותר של מניבים מאשר חלקם באוכלוסייה יניב הסתברויות רכישה גבוהות יותר מההסתברויות האמיתיות, מה שעשוי להביא להחלטות מוטעות. הפתרון הוא לעדכן את המקדמים של מודל הרגרסיה על מנת שישקפו את היחס האמיתי של מניבים ללא מניבים באוכלוסייה. במודל רגרסיה לוגיסטית הסיפור הוא פשוט יותר, שכן למעט החותך (Intercept) של משוואת הרגרסיה כל שאר המקדמים במשוואת הרגרסיה לא משתנים, מה שמחייב לעדכן רק את החותך של משוואת הרגרסיה הסופית (Ben Akiva and Lerman, 1987). במודלים אחרים העדכון מורכב יותר, אבל כמובן שאי אפשר להתעלם מתופעה זו כדי להימנע ממודלים מוטים.

4. כשלים הנובעים מהטיות בנתונים (Data bias)

נושא ההכנה של הנתונים למודל חיזוי הוא אחד משלבי העיבוד המקדים (Preprocessing) של בניית המודל. האמת שמדעני הנתונים מקדישים הרבה זמן (עד 60%–70% מזמנם ואולי יותר) להכין את הנתונים למודל, לנקות אותם, לזוג רשומות ממקורות שונים, לבצע סטנדרטיזציה של הנתונים וגם להעשיר אותם עם נתונים רלוונטיים ממקורות אחרים. הספרות העוסקת בעיבוד המקדים כוללת גם דיון בפעולות הנדרשות בהכנת הנתונים כדי למנוע כשלים בבניית מודלים של חיזוי הנובעים מהטיות בנתונים. דיון נרחב בנושא זה ניתן למצוא גם בוויקיפדיה ובאתרים מקוונים רבים. מבין הספרים שנכתבו על הנושא נזכר שניים מהם (Brownlee (2021), Jägare (2020). Mehrabi et al. (2019). רשימה של 23 הטיות אפשריות בנתונים עבור מודל חיזוי, החל בבעיות של למידת מכונה ועד בעיות של למידה עמוקה.

על מנת לשפר את הטיב והדיוק של המודלים לחיזוי. הבנה של הבעיה העסקית (*Domain knowledge*) יכולה להועיל מאוד כדי להגדיר את הפונקציות "הנכונות" עבור מודל החיזוי.

4.3 חוסר איזון בנתונים (Data imbalance)

חוסר איזון בנתונים מתייחס לתופעה הנפוצה של חוסר אחידות בנתוני הקלט על פני פלחים שונים באוכלוסיית היעד. למשל, לקוחות ותיקים שנמצאים בבסיס הנתונים כבר זמן רב ולכן צברו היסטוריה של קניות ופניות בעבר, מול לקוחות חדשים שרק הצטרפו לבסיס הנתונים ולא הספיקו עדיין ליצור לעצמם היסטוריה של רכישות. נתוני סקרים (*Surveys*) זמינים בדרך כלל רק לגבי לקוחות מגיבים ולא ללקוחות לא מגיבים. לעיתים נתונים מסוימים (למשל מחירים) מוזנים רק ללקוחות שנענו להצעת הרכישה ולא ללקוחות שלא הגיבו לפנייה אליהם, והדוגמאות הן רבות. חוסר האיזון בנתונים עשוי לעוות את תוצאות המודל. למשל, שימוש בנתונים שזמינים רק למגיבים, כגון מחיר הפריט, עשוי להניב מודל "מושלם" במובן שהמחיר של הקנייה הקודמת הוא המנבא האולטימטיבי של הקנייה הבאה, מה שכמובן לא נכון. בנייה של מודל תגובה על בסיס מדגמים לא מאוזנים יכול להניב מודל שגוי. לדוגמה, בנייה של מודל אחד על מדגם הכולל גם לקוחות ותיקים וגם לקוחות חדשים עשוי ליצור "אפליה" בין משתני הקלט בכך שהמודל "מחזק" משתנים מסבירים מסוימים ו"מחליש" משתנים אחרים. לכן יש צורך להפעיל שיקול דעת על מנת לבנות מודל תגובה מאוזן למרות חוסר האיזון בנתונים. *Kim et al. (2019)* מציעים להתמודד עם בעיות של חוסר איזון הנתונים באמצעות גישות של דגימת יתר (*Oversampling*). אפשרות אחרת היא לבנות מודל נפרד עבור לקוחות ותיקים ועבור לקוחות חדשים, או לעדכן את תוצאות המודל אפילו תוך שימוש בשיקולים אינטואיטיביים על מנת להתחשב בעיוות בנתונים, ועוד.

4.4 נתונים חסרים (Missing values)

נתונים חסרים הם תופעה נפוצה בבסיסי נתונים גדולים, במיוחד במקרים שבהם הנתונים בבסיס הנתונים מגיעים ממקורות שונים. לרוב, נתונים חסרים הם ברמה של המאפיינים

(*Quartiles*). כל מקטע כזה מיוצג באמצעות משתנה קטגורי המקבל ערך 1 אם המאפיין X עבור הלקוח נופל בתחום של המקטע, 0 - אם לא. בפונקציה הליניארית למקוטעין מחלקים את תחום ההשתנות של המשתנה המסביר למספר מקטעים ליניאריים לא חופפים קשורים ורציפים (*Non-overlapping and continuously-linked linear segments*), כל מקטע עם שיפוע משלו. צורת הקשר הלא ליניארי נקבעת על סמך האומדים של המקדמים של המשתנים הקטגוריים בפונקציית המדרגות, והאומדים של השיפועים של המקטעים הליניאריים בפונקציה הליניארית למקוטעין. שתי הפונקציות האלה מאפשרות להגדיר מנעד רחב של קשרים לא ליניאריים בין משתנים שנקבעים על ידי הנתונים עצמם ולא על סמך פונקציות "סגורות" שמוגדרות מראש.

4.2 פונקציות של משתנים מסבירים Functions of explanatory (variables)

לעיתים קרובות יכולת החיזוי של משתנה מסביר באה לידי ביטוי לא במשתנה עצמו אלא באיזושהי פונקציה של המשתנה המסביר עצמו או אפילו בשילוב עם משתנה מסביר אחר. דוגמאות טיפוסיות הן פרופורציות (*Proportions*) או יחסים (*ratios*) בין משתנים. למשל, *Tam & Kiang (1992)* השתמשו ב-19 יחסים פיננסיים מקובלים כדי לחזות כשלים של בנקים. בבעיות שיווק ישיר שיעור התגובה (*Response ratio*) המבטא את היחס בין מספר הרכישות שלקוח עשה בעבר לבין מספר הפניות שנעשו אליו בעבר הוא מנבא טוב יותר של הרכישה הבאה מאשר מספר הרכישות /או מספר הפניות. באמצעות פרופורציות ניתן גם לתקן (*Scale, standardize*) משתנים. למשל, במקום להשתמש בסך ההוצאות הכספיות על קניות מהחברה באינטרוול זמן נתון בעבר בתור משתנה מסביר, ניתן להשתמש בפרופורציה של ההוצאה הכספית על קניית מוצרים באינטרוול הזמן הזה יחסית לכלל ההוצאה הכספית בתקופה הזו. היתרון של פרופורציות הוא שהן ניתנות להשוואה. למשל, בבעיות שיווק יותר הגיוני להשוות את שיעורי התגובה של לקוחות במקום להשוות את מספר הרכישות, מכיוון שלמספר הרכישות אין משמעות אלא אם מייחסים אותן למספר הפניות שנעשו ללקוח בהצעה לקנות את המוצר.

תקצר היריעה מלפרט את טווח הפונקציות האפשריות לבניית מודלים, אבל ראוי להדגיש שלפונקציות האלה יש חשיבות רבה

4.5 חריגים (Outliers)

ערכים חריגים הם הקיצוניות הנגדית למשתנים חסרים. חריגים הם מאפיינים עם ערכים שנמצאים במרחק רב של מספר סטיות תקן מהערך הממוצע של המאפיין על פני כל הלוקוחות בקובץ האימון. כמו במקרה של משתנים חסרים, גם כאן קיימת הבעיה של שקלול תמורות. מחד, השמטה של תצפיות עם ערכים חריגים כרוכה בהפסד אינפורמציה. מאידך, שיתוף תצפיות עם ערכים חריגים בתהליך בניית המודל עשוי לעוות את תוצאות המודל. פתרון סביר לבעיית החריגים היא לקצץ (*Trim*) ערכים חריגים מלמעלה לרמה השווה לממוצע של המאפיין פלוס מספר מוגדר מראש של סטיות תקן (למשל 5), וכשמדובר על ערכים חריגים מלמטה, לקצץ את הערכים החריגים לרמה השווה לממוצע של המאפיין מינוס מספר מוגדר מראש של סטיות תקן. מספר סטיות התקן יכול להיות זהה לכל המאפיינים או שונה ממאפיין למאפיין, תלוי ביישום.

4.6 נתונים "רועשים" (Noisy data)

נתונים רועשים הם מאפיינים עם פרופורציה קטנה מאוד של תצפיות "מאובסות". למשל, פחות מ-0.5% מהתצפיות במדגם האימון מכילים ערכים עבור המאפיין. דוגמה אחרת היא משתנה בינארי, למשל מגדר, שעבורו אחוז התצפיות המכילות את הערך 1 (נשים) או 0 (גברים) הוא מאוד קטן (למשל פחות מ-0.5% מהתצפיות במדגם האימון) וההיפך, אחוז התצפיות המכילות את הערך 1 או 0 הוא מאוד גבוה (למשל למעלה מ-99.5% מהתצפיות במדגם האימון). מאפיינים אלה לא מכילים מספיק אינפורמציה כדי לשמש משתנים מסבירים במודל תגובה ולכן יש להוריד אותם מקובץ האימון.

4.7 משתנה תלוי "מזוהם" (Contaminated)

הסטיות בנתונים יכולות לקרות גם במשתנה התלוי. דוגמה טיפוסית היא משתנה תלוי שהזדהם על ידי אחד או יותר מהמשתנים המסבירים. לדוגמה, במודל תגובה בינארי עם משתנה תלוי המבטא את תגובת הלוקוחות להצעה לרכוש מוצר מסוים (1 – כן, 0 – לא), אם הערך הכספי של הקנייה הנכחית כלול במשתנה המסביר "רמת ההוצאה" על רכישת מוצרים בעבר, המשתנה הזה יכול לנבא באופן מושלם את

(משתנים מסבירים). לדוגמה, משתנה המגדר (זכר/נקבה) שחסר במספר גדול של תצפיות, משתנה המחיר שמוגדר רק עבור מניבים, ועוד. אם הפרופורציה של הנתונים החסרים עבור מאפיין מסוים היא גבוהה, הדבר עשוי לעוות את תוצאות המודל, במיוחד אם מספר המאפיינים עם נתונים חסרים הוא גדול. למעשה, יש כאן שקלול תמורות (*Tradeoff analysis*). מחד, הורדה של תצפיות ממדגם האימון עם נתונים חסרים כרוך בהפסד אינפורמציה. מאידך, שיתוף תצפיות עם נתונים חסרים בתהליך בניית המודל יניב קרוב לוודאי מודל שגוי. פשרה אפשרית היא לסנן תצפיות עם אחוז גבוה במיוחד של נתונים חסרים באחד או יותר מהמאפיינים (למשל, מעל 50% מהתצפיות הן חסרות מגדר). לגבי שאר המאפיינים עם משתנים חסרים, ניתן לבטא את ההשפעה של המשתנה החסר באמצעות הגדרה של משתנה מסביר נוסף לכל מאפיין שיקבל את ערך 1 אם קיימת לפחות תצפית אחת בקובץ האימון עם מאפיין חסר, ו-0 אחרת. משתנים אלו יצטרפו לרשימת המשתנים המסבירים הפוטנציאליים וייקחו חלק בתהליך בחירת המשתנים למודל. אפשרות אחרת היא "לשתוד" (*Impute*) ערך מספרי עבור כל משתנה חסר. איזה ערך לשתוד? זה כבר תלוי בסוג המאפיין. במקרה של משתנים קטגוריים, למשל מצב משפחתי (*Marital status*) עם ארבעה ערכים אפשריים – רווק(ה), נשוי(אה), אלמן(ה), גרוש(ה), הערך הסביר ביותר לשתוד במקום מצב משפחתי חסר הוא השכיח (*Mode*) של המשתנה, כלומר הערך של המשתנה הקטגורי הנפוץ ביותר. לדוגמה, אם רוב התצפיות במדגם האימון הן של לקוחות נשואים, נמיר מצב משפחתי חסר בערך "נשוי". במשתנים נומריים (כגון הוצאה כספית על קניות בעבר) האפשרויות הן לשתוד עבור משתנה חסר את הממוצע של המאפיין עבור כל התצפיות בקובץ האימון, החציון של המאפיין, הערך המקסימלי של המאפיין, הערך המינימלי או כל אחוזון אחר (למשל האחוזון ה-95%), ועוד. דרך מתוחכמת יותר היא למצוא את הערך המיטבי על פי קריטריון מסוים. בבעיות שיווק ישיר קריטריון אפשרי הוא שיעור התגובה (היחס בין מספר הקניות למספר הפניות). לדוגמה, עבור משתנה בינארי (נשים/גברים) נשתוד את הערך 0 (גברים) או 1 (נשים) שעבורו שיעור התגובה הוא הקרוב ביותר לשיעור התגובה בכל קובץ האימון. הספרות מציעה גם שיטות מתוחכמות יותר להמרת משתנים חסרים עבור משתנים נומריים, למשל לאמוד את הערך של המשתנה להמרה בשיטות רגרסיה.

החלטות הרכישה של הלקוחות במבצע הדיור ונותן מודל שהוא "טוב מדי מלהיות אמיתי" (*Too good to be true*). כמוכן שמודל כזה שגוי מיסודו. הפתרון לבעיה זו הוא לדאוג שהמשתנה התלוי יהיה "נקיי" וחסר מההשפעה של משתנה מסביר כלשהו.

5. כשלים הנובעים מיישום של מודל החיזוי (implementation pitfalls)

כשלים במודל החיזוי יכולים לנבוע גם מבעיות ביישום מודל החיזוי לצורך קבלת החלטות. הכשלים העיקריים נובעים בעיקר מהטיות על פני זמן בגין האופי הדינמי של בסיסי הנתונים שלא שוקטים על שמריהם ומשתנים באופן תדיר לאורך זמן. אנו נדון כאן בשתיים מההטיות האלה.

5.1 הטיות הנובעות משינויים בתמהיל הלקוחות על פני זמן (selection bias)

מבצעי שיווק הם בדרך כלל מבצעים שחוזרים על עצמם לאורך זמן. לדוגמה, נניח שרשרת של מבצעי שיווק עתידיים עם תופעת המשפך (*Funnel effect*). במבצעים אלה קהל היעד בכל מבצע שיווק נבחר על סמך תוצאות מבצעי השיווק הקודמים. כאמור, אם מדובר על מוצר או שירות חדש שלא הוצע לצרכנים בעבר, התהליך מתחיל עם דיור מבחן (*Test mailing*) שבו מציעים את המוצר למדגם אקראי מתוך אוכלוסיית היעד. על בסיס התוצאות של דיור המבחן בונים מודל תגובה שעל פיו בוחרים את הלקוחות שישתתפו במבצע השיווק הראשון בשרשרת (*Roll*). הלקוחות שישתתפו במבצע השיווק השני בשרשרת (*Reroll*) נבחרים על סמך מודל תגובה המתבסס על תוצאות מבצע השיווק הראשון (*Roll*). באופן דומה, מבצע השיווק השלישי בסדרה (*Rereroll*) נבחר על פי מודל התגובה שנבנה על בסיס האוכלוסייה שהשתתפה במבצע השיווק השני (*Reroll*), מבצע השיווק הבא בסדרה מתבסס על מודל תגובה שנבנה על בסיס קהל הלקוחות שלקחו חלק במבצע הקודם (*Rereroll*), וכך הלאה.

התהליך הזה "סובל" מהטיות בחירה (*Selection bias*) בגין שימוש במדגמים מוטים. מכור, כשמדובר על מוצר חדש, קהל היעד שמשותף במבצע השיווק הראשון (*Roll*) נבחר על סמך מודל תגובה שמתבסס על דיור המבחן, ולכן הוא כבר לא מייצג את כלל האוכלוסייה. והרי לנו מקור להטיית בחירה בתכנון מבצע השיווק השני (*Reroll*), שכן הוא מתבסס על מדגם מוטה של אוכלוסיית ה-*Roll*.

אם נעבור הלאה, בניגוד ללקוחות המשתתפים במבצע השיווק הראשון (*Roll*), שכוללים בדרך כלל לקוחות שלא נחשפו בעבר למוצר המוצע, הרי קהל היעד במבצע השיווק השני (*Reroll*) כולל לקוחות משני סוגים: כאלה שנחשפו כבר למוצר (*Exposed*), וכאלה שעדיין לא נחשפו (*Unexposed*).

הלקוחות ב-*Reroll* שכבר נחשפו למוצר (*Exposed*) הם אלה שהשתתפו במבצע השיווק הראשון (*Roll*), לא הגיבו להצעה לרכוש את המוצר המוצע, אבל עדיין שייכים לקהל היעד של ה-*Reroll* מכיוון שהם עונים על הקריטריונים עבור מבצע השיווק השני (למשל, לקוחות שביצעו לפחות קנייה אחת של מוצר אחר בשלוש השנים האחרונות).

לעומתם, הלקוחות ב-*Reroll* שלא נחשפו עדיין למוצר (*Unexposed*) הם משני סוגים:

- לקוחות חדשים שהצטרפו לבסיס הנתונים בתקופה שבין מבצע השיווק הראשון (*Roll*) למבצע השיווק השני (*Reroll*).
- לקוחות ותיקים שלא ענו על הקריטריונים להשתתפות במבצע השיווק הראשון (*Roll*) אבל הם "התבגרו" (*Graduated*) מאז וכעת עונים על הקריטריונים למבצע השיווק השני (*Reroll*). למשל, אנשים שקנו מוצר אחר מהחברה בטווח הזמן בין שני מבצעי השיווק ולכן הם זכאים להשתתף במבצע השיווק השני בתור קונים עכשוויים (*Recent buyers*).

ומכאן, בנוסף להטיות הבחירה, קהל היעד עבור מבצע השיווק השני (*Reroll*) אינו תואם (*Compatible*) את קהל היעד עבור מבצע השיווק הראשון (*Roll*), מפני שהוא מכיל סוגים "שונים" של לקוחות *Exposed* ו-*Unexposed*, בשעה שקהל היעד עבור מבצע השיווק הראשון (*Roll*) כולל רק לקוחות מסוג *Unexposed*. השאלה שעולה במקרה הזה היא איך מעדכנים את הסטבריות הרכישה של הלקוחות ב-*Reroll*

5.2 שינויים דינמיים על פני זמן (as of date)

כאמור, בסיסי נתונים של לקוחות הם דינמיים מאוד ומשתנים באופן שוטף — לקוחות חדשים נכנסים למאגר הנתונים, לקוחות ותיקים פורשים, פעילות הלקוחות משתנה על פני זמן, לקוחות אחרים מגיבים להצעות רכישה של מוצרים ושירותים, ויותר לקוחות לא נענים להצעות הרכישה. גם הפרטים הדמוגרפיים של הלקוחות משתנים לאורך זמן (למשל ניל הלקוח) וגם היסטורית הרכישה שלו (למשל סך ההוצאה הכספית על רכישות מהחברה בתקופת זמן מסוימת). מאידך, גם מבצעי השיווק נמשכים בדרך כלל לאורך זמן בעיקר בשל הצורך להצטייד במלאי מספיק של מוצרים לפני שיוצאים עם מבצע השיווק על מנת לספק את הביקוש. אחד היתרונות של מודל הרגרסיה הלוגיסטית הוא האפשרות לחזות מראש את הביקוש למוצר. כך למשל, מודל התגובה שמבוסס על תוצאות דיוור המבחן מאפשר לחזות את הביקוש למוצר התורן במבצע ה-Roll. מודל התגובה המבוסס על מבצע ה-Roll מאפשר לחזות את הביקוש למוצר התורן במבצע ה-Reroll, וכך הלאה. אבל בשל הצורך להצטייד מראש במלאי מוצרים לקראת מבצע השיווק, עוברים לעיתים מספר חודשים, אם לא יותר, בין מבצעי השיווק השונים. למשל, פרק הזמן שעובר בין דיוור המבחן לבין מבצע ה-Roll יכול לעלות על חצי שנה ויותר, וכשמדובר על מבצעי שיווק עוקבים (למשל בין ה-Roll ל-Reroll) פרק הזמן העובר עשוי להיות ארוך יותר. בתקופת הזמן האלה שבין מבצעי השיווק, בסיסי הנתונים משתנים באופן דינמי ויש לתת על כך את הדעת בבניית מודל התגובה על מנת להימנע מהטיות במודל. כאן נכנס לתמונה הנושא של *As of date*.

המטרה היא להתאים את התמהיל של הלקוחות בקבצים הלקוחים חלק בבניית מודל התגובה למועד (*As of date*) של מבצעי השיווק. למשל, נניח שדיוור המבחן נעשה בראשון בינואר 2021 ומבצע השיווק המבוסס עליו (*Roll*) משובץ לראשון באוגוסט 2021. במקרה זה התכנון הראשוני של מבצע השיווק מתבסס על מודל תגובה המבוסס על קובץ דיוור המבחן שמשקף את תמונת המצב (*Snapshot*) של בסיס הנתונים לתאריך של הראשון בינואר 2021. אבל מאחר שתמהיל הלקוחות בבסיס הנתונים בראשון באוגוסט 2021 משתנה לחלוטין, אי אפשר לבחור את הלקוחות שישתתפו במבצע השיווק (*Roll*) בפועל בהסתמך על המודל של הראשון

שכבר נחשפו למוצר המוצע, לאור העובדה שאת מודל התגובה עבור בחירת הלקוחות ל-Reroll בונים על בסיס קובץ ה-Roll שמכיל רק לקוחות שלא נחשפו עדיין למוצר?

הבעיה מסתבכת יותר במבצע השיווק השלישי (*Reroll*) שנבחר על סמך מודל תגובה המבוסס על תוצאות מבצע השיווק השני (*Roll*). כאן קהל היעד המרכיב את מבצע השיווק השני (*Roll*) כולל לקוחות שלא נחשפו עדיין למוצר המוצע (*Unexposed*) וכאלה שנחשפו (*Exposed*) למוצר פעם אחת, בשעה שקהל היעד עבור מבצע השיווק השלישי (*Reroll*) כולל לקוחות שכבר נחשפו פעמיים בעבר למוצר המוצע. ושוב מתעוררת השאלה: כיצד מעדכנים את הסתברויות הרכישה של לקוחות במבצע השיווק השלישי (*Rereoll*) שנחשפו כבר פעמיים בעבר למוצר המוצע על סמך מודל תגובה שמבוסס על קהל היעד למבצע השיווק השני (*Roll*), שכולל לקוחות שלא נחשפו למוצר המוצע וכאלה שנחשפו למוצר רק פעם אחת. וכך הלאה לגבי מבצעי שיווק נוספים בשרשרת.

שאלות אלה מסבכות מאוד את מודל התגובה עבור מבצעי השיווק הבאים בשרשרת מבצעי השיווק, ולא זה המקום לדון במודלים אלה. המטרה העיקרית של הדיון הזה היא רק "להציף" את הבעיות שמתעוררות בגין ההטיות על פני זמן על מנת לתת להם מענה כדי להימנע מהחלטות שיווקיות מוטעות.

נציין שמניסיון אמפירי, שיעורי התגובה של מבצעי שיווק החוזרים על עצמם עבור אותו מוצר הולכים ויורדים עבור כל מבצע שיווק נוסף בשרשרת השיווק. *Bucannan and Morrison* (1988) מתייחסים לאירועים אלה כתופעת ה-*"List-falloff"*. אנשי השיווק משתמשים ב"כלל אצבע" שלפיו שיעור הירידה בתגובה ממבצע השיווק הראשון (*Roll*) לשני (*Roll*) הוא כ-50% ואחר כך יורד בעוד 20% עבור כל מבצע שיווק עוקב. ברור שבמצב כזה מגיעים לנקודה שבה שיעור התגובה של הלקוחות במבצע השיווק מגיע לרמה שהופך את מבצע השיווק הבא ללא רווחי, מה שעוצר את שרשרת השיווק. אבל כמובן שמדובר כאן בכלל אצבע שלא מחליף את הצורך במודלים מדויקים יותר על מנת לקבל את ההחלטות השיווקיות הנכונות.

בינואר 2021, אלא על פי מודל מעודכן שמשקף את בסיס הנתונים במועד של הראשון באוגוסט 2021.

מכאן, אם פער הזמן בין מבצעי השיווק הוא ארוך במיוחד, ייתכן שיש צורך לבנות שני מודלים של תגובה, אחד על בסיס דיוור המבחן, על מנת לחזות את מספר ההזמנות הצפוי ורשימת הלקוחות שייקחו חלק במבצע השיווק (Roll), ומודל שני במועד מבצע השיווק (Roll) כשבוחרים את הלקוחות שייקחו חלק בפועל במבצע השיווק.

6. איחוע KDD Cup 1998

נעבור עכשיו לאירוע אמיתי המדגים את השוני בתוצאות של מודלים לחיזוי אנליטי על אף שהם מתבססים על אותם בסיסי נתונים. האירוע מתבסס על *KDD CUP*, תחרות שנתית ביוזמת מדי שנה במסגרת הכנס השנתי לנילוי ידע וכריית מידע (*KDD – Knowledge Discovery and Data Mining*). מטרת התחרות היא לעודד פיתוח כלים ומודלים חדשים לחיזוי אנליטי להתמודדות עם בעיות עסקיות וניהוליות מורכבות. התחרות מאורגנת בשיתוף עם ארגון כלשהו בתחום שמשנתה משנה לשנה, ויכול להיות עסקי, שיווקי, מדעי, רפואי ועוד. הארגון הוא זה ש"תורם" את "הבעיה התורנית" לתחרות ואת קובצי הנתונים הדרושים על מנת לתמוך בבעיה זו. הבעיה התורנית משתנה משנה לשנה ונגזרת מהתחום שבו עוסק הארגון.

במאמר זה נציג את התוצאות של תחרות ה-*KDD Cup* משנת 1998. בחרנו באירוע הזה מכיוון שהוא תואם את הדין לעיל, כלומר מדובר כאן ההצעה שיווקית למוצר יחיד בדיוור ישיר (*Solo mailing*). במקרה הזה פנייה ללקוחות לתרום לקרן צדקה (*Charity*) באמצעות טכנולוגיה של שיווק מבסיסי נתונים (*Database marketing*). הנתונים לתחרות סופקו על ידי *PVA* (*Paralyzed Veterans of America*), אחד מארגוני הנכים הגדולים ביותר בארה"ב, וכללו כ-200,000 תורמים "רדומים" שנבחרו באופן אקראי מתוך מבצע דיוור גדול לתורמים רדומים שנערך במחצית 1997 וכלל למעלה מ-3.5 מיליון לקוחות. מטרת התחרות הייתה "לעורר" (*Rejuvenate*) תורמים רדומים לקרן של *PVA* שלא היו פעילים במשך שלוש שנים ויותר ולהכניס אותם חזרה למעגל התורמים. אולם במקום לפנות למיליוני התורמים הרדומים שבחלקם המכריע לא התכוון להיענות לפנייה לתרום, הרעיון היה לבנות מודל חיזוי על בסיס מדגם

קטן יחסית של נתונים, ולהפעיל את המודל הזה על כלל בסיס הנתונים על מנת לבחור למבצע הדיוור רק לקוחות "רווחיים", כלומר לקוחות שתוחלת התרומה החזויה שלהם היא גדולה או שווה מעלות הפנייה אליהם המורכבת בעיקר מעלות הדואר והעלון הנלווה (*Brochure*), ובמקרה שלנו 0.68 דולר ללקוח. 200,000 הלקוחות הרדומים שנבחרו לתחרות חולקו באופן אקראי לשתי קבוצות שוות בגודלן: קובץ אימון, לצורך בניית המודל, שכלל פרט לכל המשתנים המסבירים גם את המשתנה התלוי — גובה התרומה בפועל של כל לקוח שהניב למבצע הדיוור; וקובץ תיקוף, לצורך תיקוף המודל, שכלל את כל המשתנים המסבירים למעט גובה התרומה בפועל של הלקוחות שהניבו למבצע הדיוור. הקבצים האלה עברו גם עיבוד מקדים על מנת להעשיר את הנתונים ולהכין את הנתונים לצורך בניית המודל.

המשתנים המסבירים הכילו יותר מ-500 משתנים משני סוגים:

- היסטוריה של תרומות לקרן במבצעי דיוור קודמים.
- משתנים דמוגרפיים שנרכשו מגורם שלישי.

כל אחד ממשותפי התחרות אמור לבנות את מודל החיזוי הספציפי שלו על בסיס קובץ האימון, ולהפעיל את מודל החיזוי שלו על קובץ התיקוף על מנת לחזות את הערך של המשתנה התלוי ולהעביר את התוצאות לוועדה המארגנת של התחרות. מאחר שהערך של המשתנה התלוי עבור קובץ התיקוף ידוע, הוועדה המארגנת יכלה לבדוק את איכות המודל של כל משתתף באמצעות השוואה של תוצאות החיזוי של המשתנה התלוי עם הערכים בפועל של המשתנה התלוי. המשתתף שהצליח לחזות "הכי טוב" את המשתנה התלוי הוא המנצח בתחרות.

בעיית ההחלטה במקרה הזה שייכת לתחום החיזוי האנליטי (*Predictive analytics*). המשתנה התלוי בבעיה זו הוא התרומה של כל לקוח לקרן (בדולרים). כלומר, מדובר כאן על מודל חיזוי עם משתנה תלוי רציף מהסוג של רגרסיה ליניארית. התפקיד של מודל החיזוי כאן הוא לאמוד את תוחלת התרומה של כל לקוח. דיוור ייחשב כרווחי אם תוחלת התרומה של הלקוח, כפי שחושבה על ידי מודל החיזוי, היא גבוהה או שווה לעלות הדיוור והעלון הנלווה (כאמור, 68 סנט לכל לקוח). פונקציית המטרה הייתה להגדיל עד כמה שאפשר את התמורה נטו לארגון הצדקה באמצעות מודלים "חכמים" של חיזוי אנליטי.

כ-50 קבוצות מהאקדמיה והתעשייה נרשמו לתחרות והורידו את קובצי האימון והתיקוף, אבל רק 21 קבוצות הגישו את תוצאות החיזוי שלהם לוועדה המארגנת. התוצאות של כל הקבוצות שהשתתפו בתחרות מפורטות בטבלאות 1-2.² טבלה 1 מציגה מספר תוצאות סטטיסטיות מסכמות, וטבלה 2 מציגה את תוצאות התחרות כשהן ממוינות בסדר יורד של הרווחיות נטו עבור קובץ התיקוף, מהקבוצה הרווחית ביותר לקבוצה הכי פחות רווחית. טבלה 2 מכילה גם את תוצאות הדיוור עבור קובץ התיקוף. קבוצה זו שימשה את הוועדה המארגנת כקבוצת הייחוס לצורך ההערכה של הביצועים של הקבוצות שהשתתפו בתחרות. חמש הקבוצות הרווחיות ביותר זוהו בשמותיהן. על מנת לא להביך את הקבוצות האחרות, הוועדה המארגנת החליטה שלא לפרסם את שמות המשתתפים מעבר לחמש הקבוצות המובילות.

קובץ התיקוף מנה 96,367 לקוחות. 4,873 לקוחות הגיבו באופן חיובי ותרמו לקרן (שיעור תגובה של כ-5%). סכום התרומות של הלקוחות המגיבים בקובץ התיקוף הסתכם ב-74,090 דולר. בניכוי עלויות הדואר והעלונים הנלווים, סך הרווח נטו משייתוף כל קובץ התיקוף במבצע הדיוור הוא 10,560 דולר. מהתבוננות בטבלה 2, קבוצת התיקוף נמצאת במקום ה-13 במדרג הקבוצות מבחינת סך התמורה נטו, כלומר 12 קבוצות הצליחו להשיג תמורה נטו גבוהה יותר לקרן הצדקה עם מספר קטן יותר של לקוחות! הקבוצה הרווחית ביותר (GainSmarts) ממליצה לכלול במבצע הדיוור רק 56,330 לקוחות, בערך מחצית מהלקוחות בקובץ התיקוף שהשתתפו בפועל במבצע הדיוור. למרות אוכלוסיית הדיוור היותר קטנה, סך התרומות נטו מהמבצע הוא 14,712 דולר, כמעט 40% יותר מאשר מבצע הדיוור לכלל אוכלוסיית התיקוף. הסיבה היא שקבוצה זו הצליחה באמצעות מודל החיזוי שלה לסנן הרבה לקוחות לא רווחיים ולהעלות את התמורה נטו. במקום השני, ובמרחק קטן מאוד מהקבוצה במקום הראשון, נמצאת חברת SAS עם 55,838 לקוחות וסך תמורה נטו של 14,662 דולר. וכך הלאה. מה שמפתיע הוא שישנן מספר קבוצות שסך התמורה נטו הייתה נמוכה מזו של קבוצת הייחוס (קובץ התיקוף), שמשמעותה הפסד כספי ל-PVA. הקבוצה במקום האחרון "הגדילה לעשות" עם תמורה נטו שלילית.

מה גרם להבדלים האלה בתוצאות בין הקבוצות השונות שהשתתפו בתחרות? אי אפשר לתת תשובה חד־משמעית

לשאלה זו בלי לרדת לפרטים של מודל החיזוי של כל קבוצה. אולם נראה שההבדלים העיקריים נובעים מהכשלים בבניית מודל חיזוי שנידונו במאמר הזה. קריטית במיוחד היא בעיית בחירת המשתנים המסבירים שנכנסו לכל מודל. נציין שמדובר כאן על בעיית חיזוי מורכבת מאוד שכוללת כ-500 משתנים מסבירים שמתוכם רק מיעוטם (אולי 30-40) הם המשתנים המשפיעים שיש להכניס למודל. הבחירה של משתנים אלה מתוך כ-500 משתנים מסבירים (בעיית ה-*Feature selection* בלשונם של אנשי כריית המידע) היא בעיה קומבינטורית מאוד מורכבת שמחייבת מיומנות רבה כדי לבנות את מודל החיזוי. בחירה לא נכונה של המשתנים המסבירים למודל החיזוי, בשילוב עם טעויות הדגימה, יכולה להביא למודל שגוי ולהחלטות עסקיות מוטעות. אולי זה מה שקרה לקבוצות בתחתית טבלה 2 שהניבו תמורה נטו קטנה מזו של קבוצת הייחוס ואף תמורה שלילית. יש לציין שגם בקרב הקבוצות ה"טובות", שככל הנראה השכילו להתמודד עם חלק מהקשיים בבנייה של מודל חיזוי, עדיין קיימת שונות גבוהה בין התוצאות עם הבדלים של כמעט 40% בסך התמורה נטו, מה שאומר שגם בקרב קבוצות אלה יש מקום לשיפור. סביר להניח שהמשתתפים בתחרות השתמשו במגוון שיטות שהוצעו בספרות המקצועית על מנת להתמודד עם בעיית בחירת המשתנים המשפיעים למודל חיזוי, שמן הסתם תרמו גם הן להבדלים בתוצאות החיזוי של הקבוצות השונות. אם נוסף לכך את שאר הכשלים בבניית מודל חיזוי שנידונו בסעיפים לעיל, זה יכול להסביר חלק מהשונות בתוצאות בטבלה 2.

המסקנה העיקרית של הדיון הזה היא שבניית מודל חיזוי בעולם נתוני העתק אינה בעיה "אוטומטית" של הפעלת תוכנת חיזוי, אלא מחייבת להתמודד עם הכשלים השונים בבניית מודלים לחיזוי. הבנה של הבעיה העסקית והכרה של התחום יכולות גם הן לתרום לבניית מודלים מדויקים יותר שמגיבים, כפי שגם עולה מטבלה 2, החלטות עסקיות טובות יותר. חשוב מאוד גם לתקף את המודל על מנת לוודא שמדובר במודל יציב ללא התאמת יתר או חסר שניתן ליישמו גם ללקוחות חדשים. במאמר הנוכחי התמקדנו רק בשיטה אחת לתיקוף המודל באמצעות השוואת תוצאות החיזוי בין קובץ האימון והתיקוף, אבל קיימות גם ואריאציות אחרות לתיקוף מודלים, כגון תיקוף צולב, תיקוף רב־ממדי כגון *10-fold validation*, ועוד. גם בדיקה אינטואיטיבית של תוצאות המודל (*Face validity*) יכולה לתרום לתיקוף המודל, אבל היא מחייבת הכרה עמוקה יותר של הבעיה (*Domain knowledge*) ובסיסי הנתונים.

2 מקור: KDD-CUP-98 Results (kdnuggets.com)

טבלה 1: נתונים סטטיסטיים (קובץ התיקוף)

Field	N	MIN	MEAN	STD	MAX	SUM
# of Responders	96,367	0	5.1%	21.9%	1	4,873
Donation Amount	96,367	\$0	\$0.79	\$4.73	\$500.00	\$76,090
Profit		-\$0.68	\$0.11	\$4.73	\$499.32	\$10,560

טבלה 2: תוצאות מפורטות (קובץ התיקוף)

Participant	N*	MIN	MEAN	STD	MAX	SUM**
GainSmarts	56,330	-\$0.68	\$0.26	\$5.57	\$499.32	\$14,712
SAS	55,838	-\$0.68	\$0.26	\$5.64	\$499.32	\$14,662
Quadstone	57,836	-\$0.68	\$0.24	\$5.66	\$499.32	\$13,954
CARRL	55,650	-\$0.68	\$0.25	\$5.61	\$499.32	\$13,825
Amdocs	51,906	-\$0.68	\$0.27	\$5.69	\$499.32	\$13,794
#6	55,830	-\$0.68	\$0.24	\$5.63	\$499.32	\$13,598
#7	60,901	-\$0.68	\$0.21	\$5.43	\$499.32	\$13,040
#8	48,304	-\$0.68	\$0.25	\$5.83	\$499.32	\$12,298
#9	56,144	-\$0.68	\$0.20	\$5.32	\$499.32	\$11,423
#10	90,976	-\$0.68	\$0.12	\$4.84	\$499.32	\$11,276
#11	62,432	-\$0.68	\$0.17	\$5.13	\$499.32	\$10,720
#12	65,286	-\$0.68	\$0.16	\$4.53	\$224.32	\$10,706
Mail entire Aud.	96,367					\$10,560
#13	64,044	-\$0.68	\$0.16	\$4.99	\$499.32	\$10,112
#14	76,994	-\$0.68	\$0.13	\$4.91	\$499.32	\$10,049
#15	54,195	-\$0.68	\$0.18	\$5.29	\$499.32	\$9,741
#16	79,294	-\$0.68	\$0.12	\$4.47	\$249.32	\$9,464
#17	51,477	-\$0.68	\$0.11	\$4.00	\$111.32	\$5,683
#18	30,539	-\$0.68	\$0.18	\$5.34	\$499.32	\$5,484
#19	50,475	-\$0.68	\$0.04	\$3.44	\$99.32	\$1,925
#20	42,270	-\$0.68	\$0.04	\$3.64	\$99.32	\$1,706
#21	1,551	-\$0.68	-\$0.03	\$3.60	\$53.32	-\$54

* N – number of responders for which the predicted donation amount > \$0.68

** SUM – sum of (Actual Donation-\$0.68) for all responders with predicted donation > \$0.68

7. סיכום

במאמר זה סקרנו מספר כשלים בבניית מודלים לחיזוי אנליטי בשלוש קטגוריות: כשלים הנובעים מהטיות בבניית מודל התגובה, כשלים הנובעים מהכנת הנתונים למודל, וכשלים הנובעים מיישום המודל לקבלת החלטות. מאחר שמודלים שונים ויישומים שונים דורשים התייחסות אחרת לנתונים, התמקדנו במאמר הזה במודלים של חיזוי אנליטי מבוססי גרסיה בתחום של שיווק ישיר. "קינחנו" את המאמר באמצעות הצגה של אירוע אמיתי, תחרות ה-*KDD CUP 1998*, שממנו ניתן ללמוד על השונות בתוצאות, על קובץ התיקוף, של מודלים שונים לחיזוי אנליטי, על אף שהם התבססו על אותם בסיסי נתונים. את התוצאות האלו השוונו לקובץ הייחוס — תוצאות הדיוור על כל קובץ התיקוף. הפרמטר להשוואה היה סך התרומות נטו לארגון הנכים (PVA). 12 קבוצות, מתוך 21 הקבוצות שהשתתפו בתחרות הניבו תוצאות טובות מאלה של קובץ הייחוס, ו-9 קבוצות הניבו תוצאות גרועות יותר. יש להניח שההבדלים בתוצאות החיזוי בין המודלים השונים נובעים במידה רבה מהאופן שבו התייחסו הקבוצות השונות לכשלים שנדונו במאמר הזה. ככל הנראה, הקבוצות שביצעוהן היו מעל אלה של קובץ הייחוס הצליחו להתגבר על חלק מהמוקשים בבנייה או ביישום של מודל התגובה, וכתוצאה מכך הצליחו לסנן את הלקוחות הבלתי רווחיים ולהעלות את סך התרומה נטו לארגון הנכים. לעומתן, הקבוצות שביצעוהן היו מתחת לקובץ הייחוס "דרכו" על מרב המוקשים בבנייה או ביישום של מודל התגובה, אם בשל חוסר ידע או בשל חוסר תשומת לב, מה שהביא לתוצאות עסקיות גרועות יותר מאלה

שהתקבלו במבצע הדיוור בפועל, ובמקרה הקיצוני ביותר אף לתמורה נטו שלילית. בשורה התחתונה, אירוע זה ממחיש עד כמה ניתן לשפר את התוצאות העסקיות באמצעות מודלים "חכמים" של חיזוי אנליטי.

המסקנה העיקרית של המאמר היא שבנייה ויישום של מודלים לחיזוי אנליטי בעולם נתוני העתק היא משימה בהחלט לא פשוטה ומחייבת היכרות טובה לא רק עם עולם המודלים לחיזוי אלא גם עם הבעיה העסקית שמנסים לפתור ועם בסיסי הנתונים הנדרשים כדי לתמוך בבעיה העסקית. במאמר זה התמקדנו בכשלים בבנייה ויישום של מודלים לחיזוי מבוססי גרסיה בבעיות של שיווק ישיר. כשלים אלה, ואחרים, קיימים לא רק במודלים של חיזוי אנליטי אלא גם בבניה ויישום של מודלים אחרים של למידת מכונה. אי אפשר במאמר אחד לסקור את כל הכשלים האפשריים. מטרת המאמר הזה הייתה להציף את העובדה שכשלים כאלה קיימים ושיש להתמודד איתם על מנת להימנע מהחלטות עסקיות שגויות. ואכן, כפי שעולה מהמאמר הזה, ההשקעה הדרושה על מנת להתמודד עם הכשלים בבניית מודל חיזוי איכותי, יציב ומשמעותי "שווה" את המאמץ, שכן היא מניבה תשואה עסקית גבוהה יותר שמתבטאת לא רק בתמורה כספית גבוהה אלא גם ביתרונות איכותיים. למשל, בבעיית שיווק, שביעות רצון גדולה יותר של הלקוחות שלא נאלצים להתמודד עם הצעות שיווקיות שאין להם כל עניין בהן.

jacobz@tauex.tau.ac.il

פרופ' יעקב זהבי

- זהבי, י., (2017), חיזוי אנליטי (*Predictive Analytics*) – הלכה למעשה. *חידושים בניהול, הפקולטה לניהול ע"ש קולר, אוניברסיטת תל אביב*, 1, 55-69.
- Akaike, H. (1974). A New Look at the Statistical Identification Model. *IEEE Trans. Auto. Control*, 19, 716-723.
- Ben-Akiva, M., and S.R. Lerman. 1987. *Discrete Choice Analysis*, the MIT Press, Cambridge, MA.
- Benjamini, Y., and Hochberg, Y. (1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of Royal Statistical Society*, 57, pp. 289-300.
- Breiman, L., Friedman, J., Olshen, R., and Stone, C. (1984). *Classification and Regression Trees*. Belmont, CA., Wadsworth.
- Brownlee, J. (2021). *Data Preparation for Machine Learning, Machine Learning Mastery*. <https://machinelearningmastery.com>
- Buchanan, B. and Morrison, D.G. (1988). A Stochastic Model of List Falloff with Implications for Repeated Mailings. *The Journal of Direct Marketing*, 2 (3), 7-15.
- DeGroot, M. H. (1993). *Probability and Statistics* 3rd edition, Addison-Wesley.
- Efoymson, M.A. (1960). *Multiple Regression Analysis in Mathematical Method for Digital computers*. Wiley, NY, pp. 191-203.
- Friedman, J., Hastie, T. and Tibshirani, R. (1998). *Additive Logistic Regression: a Statistical View of Boosting*, Technical Report, Department of Statistics, Stanford University
- Hastie, T., Tibshirani, R. and Friedman, J. (2009). *The elements of Statistical Learning*. Springer-Verlag, NY.
- Jägare, U. (2020). *Data Preparation for Dummies*. John Wiley & Sons, NJ.
- Kim, Y., Yongchan, K. and Myunghee. C. (2019). Valid Oversampling Schemes to Handle Imbalance, *Pattern Recognition Letters*, 125, 661-667.
- Li, H., Jiongcheng, L., Xiaoming. G., Binghao. L., Yuting. L. and Xinglong. I. (2019). Research on Overfitting of Deep Learning, *IEEE 15th International Conference on Computational Intelligence and Security (CIS)*.
- Lim. C., Sangwoo. H. and Jongwuk. L. (2020). *Analyzing Deep Neural Networks with Noisy Labels*, IEEE International Conference on Big Data and Smart Computing.
- Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K. and Galstyan, A. (2019). *A Survey on Bias and Fairness in Machine Learning*, <https://arxiv.org/abs/1908.09635v2>
- Miller, A. J. (2002). *Subset Selection in Regression*, Chapman and Hall.
- Schwarz, G. (1978). Estimating the Dimension of a Model. *Annals of Statistics*, 6, 486-494.
- Tam, K.Y. and Kiang, M.Y. (1992). Managerial Applications of Neural Networks: The Case of Bank Failure Predictions, *Management Science*, 38 (7), 926-947.
- Ying, X. (2019). An Overview of Overfitting and its Solutions, *Journal of Physics Conference Series*, IOP Publishing, Vol 1168. No. 2.