הפקולטה לניהול ע"ש קולר אוניברסיטת תל אביב
מאמר זה פורסם בכתב העת "חידושים בניהול", גיליון מס' 4, מאי 2019,
היוצא לאור על ידי הפקולטה לניהול על שם קולר באוניברסיטת תל אביב.

והגישה המנצחת היא...רגרסיה בצעדים

יעקב זהבי רונן מאירי

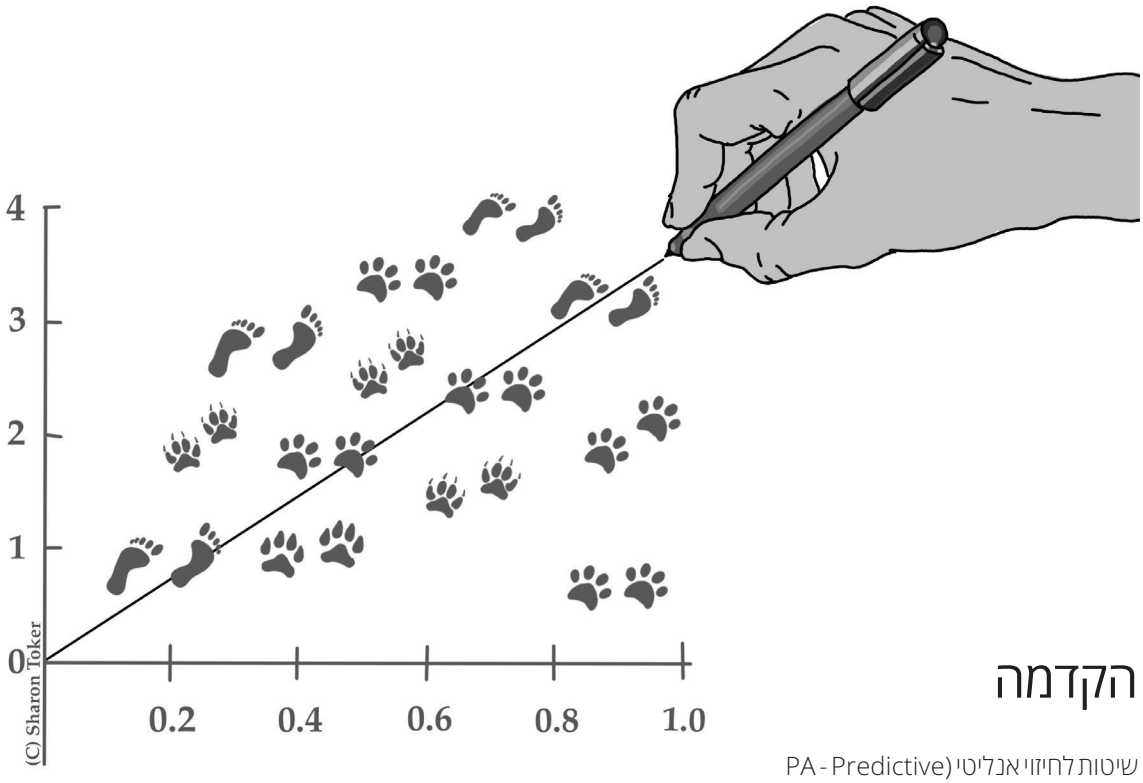
פרופ' יעקב זהבי, פרופסור אמריטוס בפקולטה לניהול ע"ש קולר באוניברסיטת תל אביב, הוא אחד מפורצי הדרך בתחום כריית המידע (Data Mining) בעולם נתוני העתק, שבו הוא מעורב במספר חזיתות – מחקר, הוראה, פיתוח תוכנה ויישומים לקבלת החלטות. פרופ' זהבי החל את הקריירה המקצועית שלו בתחום של מערכות מידע בתור מנתח מערכות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

ד"ר רונן מאירי, בעל תואר MSc בפיסיקה מהפקולטה למדעים מדויקים ותואר MBA ודוקטורט בחקר ביצועים מהפקולטה לניהול ע"ש קולר, כולם מאוניברסיטת תל אביב, משמש כיום CTO בחברת DMWay Analytics המתמחה בהנגשת עולם האנליזה המתקדמת לקהלים חדשים באמצעות למידת מכונה אוטומטית. בעבר שימש יועץ חיצוני לארגונים שונים בתחום של מדעי הנתונים (Data Science), מנהל מוצר בחברת Octavian שפיתחה מערכות Back Office לתחום הפיננסי, לרבות מערכות אנליזה מתקדמות, מדען נתונים (Data Scientist) בחברת Urban Science שפיתחה מערכת למידת מכונה אוטומטית לבניית מודלים לחיזוי, מדען נתונים בחברת IDM שפיתחה מערכות לחיזוי בעולם הביטוח, וכן מילא תפקידים נוספים בתעשיית הטכנולוגיה בארץ. בנוסף, ד"ר מאירי מרצה בתוכניות ללימודי תואר שני במנהל עסקים בפקולטה לניהול ע"ש קולר באוניברסיטת תל אביב.

תקציר

במחקר זה חיפשנו את הגישה המועדפת לבחירת משתנים מסבירים בעולם נתוני העתק בצורה אוטומטית עבור מודל חיזוי המבוסס על רגרסיה לינארית. התמקדנו בשלוש גישות שונות לבחירה של משתנים מסבירים למודל – גישות סטטיסטיות, גישות חיפוש סטוכסטיות וגישות של צמצום ממדים. בכל גישה בחרנו מודל מייצג אחד או שניים, ובסך הכול 4 מודלים. עבור כל מודל מצאנו את הקונפיגורציה האופטימלית והשווינו את המודלים המיטביים שהתקבלו זה לזה על מספר קובצי נתונים מתחום השיווק. תוצאות המחקר מעידות שלפחות בתחום השיווק, הגישה המובילה לבחירת משתנים מסבירים למודל חיזוי היא גישת הרגרסיה בצעדים. אנחנו מייחסים תכונה זו לעובדה שצרכנים בעולם השיווק מתנהגים בצורה עקבית ורציונלית, מה שמביא לכך שבסיסי נתונים בשיווק מתנהגים היטב (well-behaved) ולכן גם שיטה מיופית כמו רגרסיה בצעדים מצליחה לאתר את המשתנים המסבירים למודל. נציין שיש צורך במחקר נוסף כדי לבדוק אם ניתן להכליל את המסקנות האלה לתחומים אחרים. מכל מקום, נראה שבתחום השיווק הגישה המנצחת לבחירת משתנים למודל חיזוי היא... רגרסיה בצעדים.

המאמר מתבסס בחלקו על עבודת הדוקטורט של ד"ר רונן מאירי בפקולטה לניהול בהנחייתו של פרופ' יעקב זהבי.



הקדמה

שיטות לחיזוי אנליטי (PA - Predictive Analytics), אחד התחומים החמים כיום במדעי הנתונים (Data Science), עוסקות בחיזוי שיעורי התגובה לאירועים עתידיים על סמך תצפיות מהעבר שערכי התגובה עבורן ידועים. הדוגמאות לכך הן רבות: האם הלקוח יגיב להצעה שיווקית לרכוש מוצר או שירות, האם יגיש תביעה לביטוח ואם כן מה גודלה, האם החולה יתאושש מניתוח שעבר וכמה זמן יצטרך לשהות בבית החולים עד שישוחרר, מהי רמת ההוצאה הצפויה של לקוח על קניות מקטלוג, מה גודל התרומה של הלקוח לאגודה למלחמה בסרטן ועוד. האירוע שאנו רוצים לחזות הוא משתנה התגובה או המשתנה התלוי, והמשתנים שבאמצעותם אנו מנסים להסביר את האירוע הם המשתנים המסבירים המכונים גם משתנים בלתי תלויים או פרדיקטורים. אנשי כריית מידע מתייחסים אליהם כאל מאפיינים (features) (במאמר זה נתייחס למושגים מאפיינים, משתנים מסבירים, פרדיקטורים ומשתנים בלתי תלויים בתור מושגים נרדפים). מה שמאפיין את שיטות החיזוי בעולם של נתוני העתק (Big Data) הוא המספר הרב של המשתנים המסבירים שיכול להגיע למאות, ואולי גם לאלפים, של משתנים פוטנציאליים, וזאת בשונה מבעיות בתחומים אחרים, למשל כלכלה, שבהם מספר המשתנים המסבירים הוא קטן יחסית. ואמנם הבעיה העיקרית בבנייה של מודל חיזוי בעולם נתוני העתק היא לבחור את המשתנים המשפיעים ביותר על משתנה התגובה, בדרך כלל רק קומץ מהם, מתוך האוסף הגדול של המשתנים המסבירים הפוטנציאליים, תוך סינון רעשים, משתנים מיותרים ומשתנים לא רלבנטיים, וטיפול בחריגים, במשתנים חסרים, בקולינאריות ועוד. בעיה זו ידועה בתחום כריית המידע בתור הבעיה של בחירת מאפיינים (feature selection). הסטטיסטיקאים מכנים אותה בעיית ספסיפיקציה (Miller, 2002). המטרה היא לבחור את המשתנים שמשיאים פונקציה של טיב ההתאמה (goodness of fit) שנבחרה מראש, כפוף לאילוצים למניעת התאמת יתר. יש כאן שתי מגמות מנוגדות – רמת הדיוק של התחזית המיוצגת על ידי השונות של שגיאת התחזית ורמת ההטיה (bias) של התחזית כאשר מיישמים אותה על תצפיות חדשות. ככל שמוסיפים משתנים מסבירים למודל החיזוי, רמת הדיוק של המודל משתפרת אבל רמת ההטיה עולה. לכן יש צורך בשקלול תמורות (trade-off analysis) בין רמת הדיוק לרמת ההטיה של המודל כדי למצוא את קבוצת המשתנים הטובה ביותר המסבירה את התופעה שאותה רוצים לחזות (Geman et al., 1999).

מבחינה מתמטית, בעיית בחירת המאפיינים היא בעיית אופטימיזציה קומבינטורית. נסמן את מספר המשתנים המסבירים ב–K. גם אם נניח רק 2 מצבים לכל משתנה – או שהמשתנה נכנס למודל החיזוי או שהוא יוצא ממנו - מספר הקומבינציות להכנסה והוצאה של משתנים למודל מגיע ל–2^K, שהוא מספר ענק, גם אם K הוא מספר קטן יחסית, לא כל שכן אם K הוא בסדר גודל של מאות ואפילו אלפים של משתנים כפי שמקובל בעולם נתוני העתק. הבעיה מסתבכת יותר אם מתחשבים בשיקולים נוספים כגון טרנספורמציות של משתנים, אינטראקציות בין משתנים, קולינאריות, משתנים חסרים, חריגים ועוד. מכאן שבעיית בחירת המאפיינים בעולם נתוני העתק היא בעיה מורכבת מאוד שאי אפשר לפתור אותה באמצעות גישות אנליטיות אלא רק בגישות יוריסטיות.

הסיבוכיות והממדים של בעיית החיזוי מקשים על השימוש במודלים של חיזוי לצורך קבלת החלטות בעולם נתוני העתק ומגבילים אותו רק לשחקנים ולחברות גדולות שיש להם משאבים ומומחים המסוגלים להתמודד עם הקושי בבניית מודלים של חיזוי. לכן המגמה כיום בתחום האנליטיקה היא להסיר חסמים כדי להנגיש טכנולוגיות אלה לציבור הרחב. אנו עוסקים כאן ב"דמוקרטיזציה" של תהליך בניית מודלים מורכבים שתאפשר אפילו לבתי עסק קטנים להשתמש בכלים אלה לצורך קידום הפעילות העסקית שלהם[1].

אין ספק שתנאי הכרחי להנגשת מודלים של חיזוי לכול הוא אוטומציה של תהליך בחירת המאפיינים עבור בעיות חיזוי גדולות עם מספר גדול של משתנים מסבירים פוטנציאליים. ואמנם לאחרונה הוצעו מספר שיטות לבחירה של מאפיינים בבעיות חיזוי גדולות בצורה אוטומטית. במאמר זה אנו בוחנים שלוש גישות חלופיות לפתרון בעיית בחירת המאפיינים במודל רגרסיה לינארית כדי למצוא את הגישה הטובה ביותר לבחירת המשתנים המסבירים. בחרנו במודל רגרסיה לינארית לא רק משום שזהו מודל החיזוי הוותיק והנפוץ ביותר עבור בעיות חיזוי, אלא משום שמודל הרגרסיה הלינארית משמש בסיס להרבה מודלים אחרים בתחום של חיזוי. למשל, ניתן להראות שאפשר לפתור בעיות של רגרסיה לוגיסטית שבה המשתנה התלוי הוא בינרי 0/1 (כן/לא) על ידי פתרון סדרה של בעיות רגרסיה לינארית עם משקלות (Hastie et al., 2009). מה שנותן יתרון נוסף למודל הרגרסיה הוא פשטות השימוש במודל והזמינות שלו. כמו כן, מחקרים אחרים הוכיחו שלפחות בתחום השיווק, מודלים של חיזוי המבוססים על רגרסיה מספקים פתרונות שאינם נחותים מאלה שמתבססים על מודלים מתוחכמים יותר המבוססים על רשתות נוירונים (Levin and Zahavi, 1997).

מספר גישות הוצעו בספרות לתקוף את סוגיית בחירת המאפיינים בצורה אוטומטית, ובהן גישות סטטיסטיות (statistical methods), גישות חיפוש סטוכסטיות (stochastic search methods) וגישות של צמצום ממדים (dimensionality reduction methods). לצורך ההשוואה בין הגישות האלה בחרנו במודל מייצג מכל גישה: מודל רגרסיה בצעדים SWR (StepWise Regression) עבור הגישה הסטטיסטית; SA (Simulated Annealing) עבור הגישות הסטוכסטיות; PCA (Principal Component Analysis) ו-RBF (Radial Basis Functions) עבור הגישות לצמצום ממדים. את ההשוואה ביצענו על מספר בסיסי נתונים מתחום השיווק. פונקציית המטרה היא למצוא מודל שמצד אחד נותן את ההתאמה הטובה ביותר לנתונים ומצד שני הוא מודל יציב ללא התאמת יתר שניתן להכליל אותו גם לנתונים חדשים. לצורך בדיקת היציבות של המודל מקובל לחלק את אוכלוסיית המדגם לשתי תת-אוכלוסיות – מדגם האימון לצורך בניית המודל, ומדגם האימות, או התיקוף, לצורך התיקוף של המודל. כדי להביא את כל המודלים לבסיס שווה לצורך השוואה, מצאנו את הקונפיגורציה האופטימלית של הפרמטרים עבור כל אחד מ-4 המודלים הנ"ל ואז השווינו את המודלים המיטביים זה לזה. בשלב הסופי בחרנו במקדם הקביעה (coefficient of determination), R^2, לצורך ההשוואה בין המודלים, כפוף לאילוץ שהיחס בין R^2 עבור מדגם האימון ומדגם התיקוף הוא קרוב ל–1, וזאת במטרה להבטיח שנקבל מודל יציב ללא התאמת יתר.

שיטות סטטיסטיות

שיטות סטטיסטיות לבחירה של משתנים למודל מתבססת על סדרה של בדיקת השערות על המשתנים המסבירים.

1 ראו, למשל, https://hbr.org/2018/07/the-democratization-of-data-science.

השערת האפס היא שהמשתנה הנבדק אינו תורם באופן מובהק להסבר של משתנה התגובה ולכן יש להסיר אותו מהמודל. ההשערה החלופית היא שהמשתנה הנבדק משפר באופן משמעותי את טיב ההתאמה ולכן יש לכלול אותו במודל. התרומה של כל משתנה למודל מיוצגת על ידי המקדם של המשתנה במשוואת הרגרסיה. אם ההסתברות שהמשתנה אינו תורם למודל נמוכה מסף מסוים המוגדר מראש וידוע בשם רמת המובהקות - בדרך כלל 5% או פחות - אנו מסיקים שתרומת המשתנה למודל היא משמעותית ומכניסים אותו למודל.

רגרסיה בצעדים (SWR – StepWise Regression)

השיטות המסורתיות בודקות את ההשערות לגבי כל משתנה בנפרד ומוצאות את המשתנים המסבירים שיש להכניס למודל בגישה של ניסוי וטעייה. כאשר מספר המשתנים קטן (10 או פחות), מספר הקומבינציות האפשריות של בחירת משתנים למודל אינו גבוה וניתן לבחור את המשתנים המסבירים למודל בשיטות ידניות. אך כאשר מספר המשתנים המסבירים גדול, כמו בבעיות חיזוי גדולות, יש צורך להשתמש בשיטות אוטומטיות לבחור את המשתנים המשפיעים למודל. גישת הרגרסיה בצעדים StepWise Regression (SWR) היא הגישה הנפוצה לבחירה של משתנים מסבירים במודל רגרסיה רב-ממדי (Efroymson, 1960). שיטות SWR הן נוחות מאוד לשימוש ונמצאות בכל החבילות הסטטיסטיות המובילות, בכללן SAS, SPSS, R ואחרות. קיימות מספר ואריאציות לגישת SWR: גישה קדמית, גישה אחורית וגישה משולבת קדמית/אחורית. בגישה הקדמית מתחילים עם אפס משתנים במודל ומוסיפים לו משתנים מובהקים בזה אחר זה בתהליך רב-שלבי עד שמתקיימים תנאי הסיום. בגישה האחורית מתחילים עם כל המשתנים במודל ומסירים ממנו בזה אחר זה משתנים לא מובהקים עד שמתקיימים תנאי הסיום. ולבסוף, בגישה הקדמית/אחורית, שהיא המתקדמת ביותר, מוסיפים וגורעים משתנים מסבירים לסירוגין עד שמתקיימים תנאי העצירה.

גישות SWR הן גישות רב-שלביות, כאשר בכל שלב מפעילים תהליך של בדיקת השערות על סדרה של משתנים מסבירים כדי לקבוע איזה משתנה להוסיף או להסיר מהמודל. למשל, בגישה הקדמית מפעילים בכל שלב את תהליך בדיקת ההשערות על כל אחד מהמשתנים שעדיין נותרו מחוץ למודל, ומכניסים אליו את המשתנה המובהק ביותר שתורם הכי הרבה לטיב ההתאמה של המודל. בגישה האחורית מפעילים את תהליך בדיקת ההשערות על כל המשתנים שעדיין נמצאים במודל ומורידים את המשתנה שתורם הכי פחות לטיב ההתאמה שלו. בגישה הקדמית/אחורית מפעילים את התהליך לסירוגין על סדרת המשתנים שנמצאת מחוץ למודל ואת סדרת המשתנים שנמצאת בתוך המודל כדי להחליט איזה משתנים להוסיף ואיזה להסיר מהמודל.

כל גישות SWR הן גישות per-comparison, היינו הן מכניסות או מוציאות משתנים בגישה פרטנית, אחד אחד (Tukey, 1977). החיסרון העיקרי של שיטה זו שהיא נוטה להכניס יותר מדי משתנים למודל, גם כאלה שאינם מובהקים. ידוע שאם משתמשים ברמת מובהקות α של 5% בגישות per-comparison עם השערה יחידה, אחד מכל 20 משתנים, אפילו כאלה שאין להם קשר למשתנה התלוי, ייכנס למודל, בממוצע, באופן שגוי. שגיאה זו ידועה בתור טעות מסוג ראשון (Type-I error). לעומת זאת, כשמדובר בבדיקת השערות מרובות, ההסתברות לטעות מסוג ראשון לגבי משתנה יחיד עולה אף היא. ניתן להראות, תוך שימוש בשיקולים הסתברותיים פשוטים, שבגישות per-comparison הכוללות k משתנים, ההסתברות לבצע לפחות טעות מסוג ראשון אחת (שמשמעותה להכניס משתנה לא מובהק למודל) היא $1 - (1-\alpha)^k$, הגבוהה בהרבה מרמת המובהקות המקובלת. למשל, עבור α=5% ו-k=10 ההסתברות לטעות מסוג ראשון קרובה ל-40%(!), עובדה שמגדילה את מספר המשתנים הלא מובהקים במודל. ניתן להקטין את הטעות מסוג ראשון בפועל על ידי הקטנת רמת המובהקות מתחת ל-5%, מה שמגדיל הטעות מסוג שני (Type-II error) שמשמעותה להסיר משתנה מובהק מהמודל. למרבה הצער, אי אפשר להקטין בו זמנית את הטעות מסוג ראשון ואת הטעות מסוג שני. למעשה, הקטנת הטעות מסוג ראשון מגדילה את הטעות מסוג שני ולהפך. לכן יש צורך בשיקלול תמורות כדי למצוא את המשתנים שיש להכניס למודל. להלן נדון בקצרה בשתי הגישות המובילות שהוצעו בספרות לעדכון רמת המובהקות במטרה להקטין את הסיכוי לטעויות - גישת Bonferroni וגישת False Discovery Rate (FDR).

התיקון על פי בונפרוני

גישת בונפרוני (Bonferroni) מטפלת בהסתברות לבצע שגיאה מסוג ראשון במשפחה של מבחני השערות מהסוג של per-comparison (Hochberg and Tamhane, 1987). בגישה זו רמת המובהקות α מתעדכנת לרמה של $\alpha^* = \alpha/k$ כאשר k הוא מספר המשתנים הלוקחים חלק בתהליך של בדיקת ההשערות, כמתואר לעיל, כדי לבחור את המשתנה שיש להוסיף או להסיר מהמודל. לדוגמה, במודל רגרסיה שעדיין נותרו בו 10 משתנים שיש לבדוק איזה מהם ייכנס למודל, וברמת מובהקות של 5%, רמת המובהקות על פי בונפרוני היא 0.5% (5%/10=). רמת מובהקות זו משתנה משלב לשלב בהתאם למספר מבחני ההשערות בכל שלב. ה"מהדרין" מחליפים את k במספר המשתנים המסבירים K, מה שמקטין את רמת המובקות על פי בונפרוני עוד יותר. הקטנת רמת המובהקות מקטינה את הטעות מסוג ראשון כיוון שהיא מבטיחה שרק משתנים מובהקים מאוד ייכנסו למודל. אבל כשעוסקים במספר משתנים גדול, העדכון על פי בונפרוני מקטין את הסיכוי לטעות מסוג ראשון (רמת המובהקות) לרמה כה נמוכה שהיא מונעת מהרבה משתנים טובים מלהיכנס למודל ומגדילה את הטעות מסוג שני.

שיעור התגליות השגויות (FDR – False Discovery Rate)

שיטות per-comparison הן מקלות מדי ועשויות להכניס יותר מדי משתנים למודל ובכך להגדיל את הטעות מסוג ראשון. לעומתן, גישות בונפרוני הן נוקשות מדי ועשויות למנוע ממשתנים מובהקים מלהיכנס למודל ובכך להגדיל את הטעות מסוג שני. שיטת FDR – שיעור התגליות השגויות - מציעה פשרה המעדכנת את הטעות מסוג ראשון α עבור כל משתנה בתהליך בדיקת ההשערות כפונקציה של מספר המשתנים שנכנסו למודל (Benjamini and Hochberg, 1995; Benjamini and Yekutieli, 2001). נציין שהמינוחים של שיטת FDR לקוחים מהתחום של תגליות מדעיות (למשל, ברפואה) שבהן מעמתים את ההשערה שאין תגלית (השערת האפס) כנגד ההשערה שיש תגלית (ההשערה החלופית). מכאן המושג תגלית שגויה המתייחס למצב שבו דוחים את השערת האפס (בהסתברות α) שאין תגלית כאשר היא נכונה, ומקבלים את ההשערה החלופית שיש תגלית אף על פי שאינה נכונה. על פי הרעיון שמאחורי שיטת FDR, ככל שמתקבלות יותר תגליות (נכנסים יותר משתנים למודל) יש לנקוט גישה פחות שמרנית כלפי הכנסת משתנים נוספים למודל באמצעות הגדלת רמת המובהקות, מה שמוריד את הסיכוי לטעות מסוג שני. אבל כאשר הטעות מסוג שני יורדת, הדבר מאפשר ליותר משתנים מובהקים להיכנס למודל. יתרונה של שיטת FDR שהיא מביאה לאיזון בין הטעות מסוג ראשון לטעות מסוג שני ועל כן מעלה את הסבירות לקבל מודל טוב יותר.

נציין שהרעיון של שימוש במודל SWR עם שיטת FDR שונה במעט מגישת FDR המקורית של Benjamini ו–Hochberg (1995). גישת FDR המקורית בודקת את המובהקות של כל המשתנים המסבירים במודל הרגרסיה באופן סימולטני ואז משתמשת ברמות המובהקות המחושבות על פי גישת FDR כדי לבחור משתנים למודל. שיטה זו מתאימה לאוסף של משתנים מסבירים ללא קורלציה ביניהם או עם קורלציה מועטה, דבר שאינו נכון לגבי מודלים רב-ממדיים של רגרסיה בעולם העסקי, ובוודאי כשמדובר במספר גדול מאוד של משתנים מסבירים. לכן מתבקש כאן השילוב של מודל SWR, שנועד להתמודד עם בעיית הקולינאריות בין משתנים, עם גישת FDR שמאזנת בין הטעויות מסוג ראשון ושני, ובכך ליהנות מכל העולמות – גם לטפל בבעיית הקוליניאריות וגם לשלוט בהסתברויות לטעויות. שילוב זה חייב אותנו לשנות את מודל SWR המקורי על מנת לשנות את רמות המובהקות, בכל איטרציה, בהתאם למספר המשתנים שכבר נכנסו למודל, כפי שמתחייב מגישת FDR. זהו האלגוריתם שבו השתמשנו להלן לצורך המחקר המשווה בין הגישות השונות לבחירת משתנים למודל הרגרסיה.

שיטות חיפוש סטוכסטיות (Stochastic Search Methods)

שיטות חיפוש סטוכסטיות משתמשות בתהליך חיפוש אקראי, ועם זאת שיטתי, למציאת הפתרון האופטימלי לבעיית אופטימיזציה. בניגוד לשיטות SWR שמחפשות בכל שלב פתרון המשפר את הערך של פונקציית המטרה, גישות סטוכסטיות מקבלות גם פתרונות שאינן משפרות בהכרח את הערך של פונקציית המטרה אך מסיטות את תהליך החיפוש לאזור אחר במרחב הפתרונות, עובדה המעלה

את הסיכוי שהפתרון הסופי יתכנס לאופטימום גלובלי (במקום לאופטימום לוקלי). שיטות מובילות הן Simulated Annealing או בקיצור SA (van Laarhoven et. al., 1987), אלגוריתמים גנטיים (Genetic Algorithms) ובקיצור GA (Goldberg, 1989), ואחרות. במאמר זה התמקדנו בגישת SA כדי לתקוף את בעיית בחירת המאפיינים למודל חיזוי.

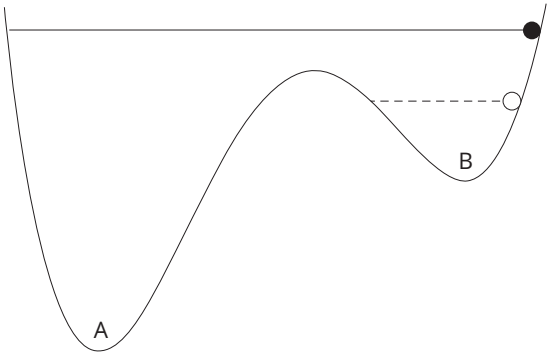
Simulated Annealing - SA

Annealing הוא תהליך של קירור מערכות פיסיקליות במטרה לשנות את התכונות האנרגטיות או המכניות של המערכת. לדוגמה, קירור גוף מתכת או גוף זכוכית כדי לשנות אותו ממצב נוזל למצב מוצק. מערכת פיסיקלית מורכבת ממספר רב של חלקיקים. פיזור האנרגיה של החלקיקים במערכת פיסיקלית תלוי בטמפרטורת המערכת והוא מתפלג בהתפלגות בולצמן (Boltzmann). ככל שטמפרטורת המערכת גבוהה יותר, נקבל ממוצע אנרגיה גבוה יותר ופיזור רחב יותר של כמות האנרגיה של כל חלקיק במערכת.

אפשר לתאר את פונקציית האנרגיה כמעין קערה בעלת מספר נקודות מינימום. חלקיק שרמת האנרגיה שלו נמוכה ינוע בתוך הקערה המקומית שבה הוא נמצא אך אף פעם לא יצליח לצאת ממנה (ראו איור 1). חלקיק שרמת האנרגיה שלו גבוהה דיה יצליח לעבור מעל גובה דופנות הקערה המקומית למקום אחר במרחב. במצב סטטי, חלקיקי המערכת מתנגשים זה בזה ומעבירים אנרגיה מאחד לשני כך שאנרגיית כל חלקיק יכולה להשתנות ולכל חלקיק יש סיכוי (לפעמים קלוש) להצליח להגיע לרמת אנרגיה גבוהה במידה מספקת כדי לעבור מעל המכשול סביב המינימום המקומי.

אם יוצאים מהמצב הסטטי ומקררים את המערכת, מפחיתים את האנרגיה הממוצעת של החלקיקים. כאשר תהליך הקירור מספיק איטי, המערכת תהיה בשיווי משקל (שמשמעותו שהתכונות הפיסיקליות הממוצעות שלה נשמרות קבוע) בכל שלב של תהליך הקירור. אם נמשיך את תהליך הקירור בצורה הדרגתית אך איטית, נצפה שכל החלקיקים יגיעו למינימום הגלובלי של המערכת כאשר הטמפרטורה תשאף ל-0. לעומת זאת, כאשר תהליך הקירור מהיר, המערכת לא תגיע לשיווי משקל וכל חלקיק ייתקע במינימום המקומי שלו.

איור 1: פונקציית האנרגיה



באיור 1 מוצגים שני חלקיקים שלכל אחד רמת אנרגיה שונה. ציר y מציין את רמת האנרגיה של החלקיק וציר x את מיקום החלקיק. אנרגיית החלקיק שעל קו הגובה המקווקו נמוכה מדופן הקערה שנקודת המינימום המקומי שלה היא B ולכן הוא יהיה לכוד בחלק הימני של הגרף. לעומת זאת, לחלקיק שעל קו הגובה הרציף יש אנרגיה גבוהה מספיק כדי להתחמק מהמינימום המקומי B ולהימצא במרחב המוגדר ע"י המינימום A.

Simulated Annealing הוא אלגוריתם לחיפוש המינימום (או המקסימום) הגלובלי של בעיות אופטימיזציה המחקה את תהליך החימום והקירור של מערכות פיסיקליות שתואר לעיל.

אלגוריתם SA הוא תהליך רב-שלבי. בכל שלב של התהליך יש שתי אפשרויות לשינוי מצב המערכת: שינוי בכיוון המינימום המקומי ושינוי נגד הכיוון, גם אם הערך המתקבל הוא פחות טוב. כדי לבחור את הכיוון הרצוי משתמשים בהתפלגות בולצמן בטמפרטורה t כך שתמיד קיים סיכוי, בהסתברות מסוימת, לנוע בכיוון המנוגד לאופטימום של המערכת ולדלג מעל ה"בור" של האופטימום המקומי. ככל שהטמפרטורה נמוכה יותר (ושואפת ל-0) הסיכוי לנוע נגד הכיוון יפחת, ואם התהליך מבוצע באיטיות מספקת הוא יתכנס לאופטימום הגלובלי.

נציין שתהליך בחירת המאפיינים למודל חיזוי הוא למעשה בעיית אופטימיזציה שבה אנחנו מנסים למצוא אופטימום של פונקציית מטרה, במקרה שלנו את הפונקציה של טיב ההתאמה. טבלה 1 מפרטת את הפרמטרים המקבילים של גישת SA בבעיה פיסיקלית ובבעיית אופטימיזציה.

טבלה 1: הקבלה בין הפרמטרים של שיטת SA לפרמטרים של בעיית אופטימיזציה

SA	תהליך אופטימיזציה
אנרגיה	פונקציית המטרה - טיב ההתאמה
מצב המערכת (state)	פתרון אפשרי
טמפרטורה	משתני הבקרה
שווי משקל תרמי	ממוצע הערך של פונקציית המטרה אינו משתנה

יישום האלגוריתם לפתרון בעיית בחירת המאפיינים מורכב מהשלבים הבאים:

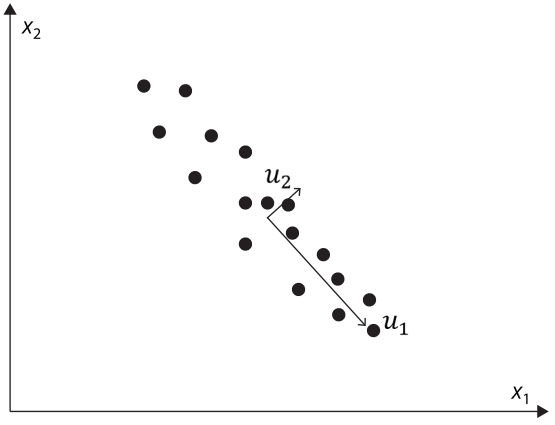
- אתחול התהליך עם פתרון ראשוני (state) של בעיית בחירת המשתנים (בדרך כלל באופן מקרי) ובחירה של קבוצת משתני הבקרה (טמפרטורה).
- קבלה של הפתרון הנוכחי אם הוא משפר את פונקציית המטרה (energy). אחרת, מקבלים את הפתרון בהסתברות p המחושבת על סמך התפלגות בולצמן, על אף שפתרון זה אינו משפר את פונקציית המטרה, וזאת כדי לתת סיכוי לתהליך SA להיחלץ מנקודת האופימום הלוקלי.
- יצירה של פתרונות שכנים לפתרון הנוכחי באמצעות הוספה או הסרה של אחד המשתנים עד קבלת פתרון שיווי משקל תרמי. במקרה שלנו - ערך פונקציית המטרה נשאר פחות או יותר קבוע עבור סדרה של פתרונות שכנים.
- קירור – כאשר מקבלים פתרון שיווי משקל, ממשיכים עם הפתרון האחרון בסדרה, משנים את הערכים של משתני הבקרה וחוזרים לשלב הקבלה.
- התהליך מסתיים כאשר מתקיימים תנאי הסיום (לדוגמה, מספר איטרציות שנקבע מראש או כאשר השינוי בערך של פונקציית המטרה בין שלב לשלב הוא מזערי).

שיטות לצמצום ממדים (PCA, RBF)

PCA – Principal Component Analysis

כאמור, בעולם נתוני העתק מספר המשתנים המסבירים הפוטנציאליים בבעיות חיזוי הוא גדול ויכול להגיע למאות משתנים, אם לא למעלה מכך. אולם חלק גדול ממשתנים אלה מתואמים ביניהם בצורה לינארית, מה שגורם לתופעה של יתירות במשתנים, שכן לפחות חלק מהמידע המוכל במשתנה אחד מוכל גם במשתנה אחר המתואם אתו. שיטת PCA (Principal Component Analysis) מתמודדת עם בעיית הקולינאריות באמצעות הגדרת מספר קטן יותר של "משתני על" לא מתואמים שמספרם קטן משמעותית ממספר המשתנים המקוריים המשמשים בתור המשתנים המסבירים במודל החיזוי במקום המשתנים המקוריים. שיטת PCA מסובבת את הצירים (הקואורדינטות) של מרחב המשתנים המקוריים לסט חדש של צירים שניצבים זה לזה (אורתוגונליים) ומגדירה סדרה חדשה של וקטורים המכונים principal components או בקיצור PC, כך שה-PC הראשון נמצא בכיוון של מקסימום השונות של מרחב המשתנים המקורי, ה-PC השני ניצב ל-PC הראשון ונמצא בכיוון של השונות המקסימלית של תת-המרחב שנותר וכך הלאה; ה-PC ה- $(m+1)^{th}$ ניצב לכל הקודמים ונמצא בכיוון של מקסימום השונות של תת-המרחב שנותר (Dunteman, 1994). איור 2 מדגים את השיטה הזו במרחב הדו-ממדי.

איור 2: תיאור סכמטי של גישת PCA במרחב הדו-ממדי



באיור 2, u_1, ה-PC הראשון, מיוצג על ידי הווקטור הנמצא בכיוון למקסימום השונות של המרחב הדו-ממדי, u_2 , ה-PC השני, מיוצג על ידי וקטור הניצב ל-PC הראשון ונמצא בכיוון של מקסימום השונות של תת-המרחב שנותר.

מבחינה מתמטית, ה-PC's הם הווקטורים העצמיים (eigenvectors) של מטריצת השונות המשותפת של

המשתנים המסבירים המקוריים לאחר שעברו תהליך של סטנדרטיזציה (כלומר, כל משתנה הוא בעל ממוצע 0 ושונות 1), והשונות של תת-המרחב המוגדר על ידי כל PC הוא הערך העצמי שלו. מטריצת השונות המשותפת היא מטריצה ריבועית סימטרית שבה כל שורה וכל עמודה מתייחסות לאחד מהמשתנים המסבירים. כל איבר על האלכסון הראשי של המטריצה מייצג את השונות של המשתנה וכל איבר שאינו נמצא על האלכסון מייצג את השונות המשותפת של המשתנים. למשל, האיבר במקום ה – (1,1) מייצג את השונות של x_1, האיבר במקום ה – (1,2) את השונות המשותפת של המשתנים x_1, x_2 וכך הלאה.

כדוגמה, נניח בעיית חיזוי עם 3 ממדים שבה מטריצת השונות המשותפת עבור המשתנים המסבירים, לאחר שעברו את תהליך הסטנדרטיזציה, מיוצגת על ידי המטריצה הבאה:

$$\mathbf{\Sigma} = \begin{bmatrix} 1.000 & 0.562 & 0.704 \\ 0.562 & 1.000 & 0.304 \\ 1.00 & 0.304 & 1.000 \end{bmatrix}$$

שלושת ה-PC's (הווקטורים העצמיים) הם:

$$u_1 = \begin{bmatrix} 0.65 \\ 0.51 \\ 0.57 \end{bmatrix} \quad u_2 = \begin{bmatrix} 0.09 \\ 0.80 \\ -0.59 \end{bmatrix} \quad u_2 = \begin{bmatrix} -0.76 \\ 0.33 \\ 0.56 \end{bmatrix}$$

השונויות של תתי-המרחב המוגדרות על ידי ה-PC's (הערכים העצמיים), שמסומנים בדרך כלל באות λ , הן: λ_1=2.05, λ_2=0.72, λ_3=0.23, בהתאמה.

נשים לב שסכום כל השונויות (הסטנדרטיות) הוא 3. מתוכן, השונות המיוצגת על ידי ה-PC הראשון מהווה 68% של השונות הכוללת (=2.05/3), ולכן ניתן ליצג את המרחב התלת-ממדי של בעיית החיזוי באמצעות ה-PC הראשון ולהקטין את הבעיה משלושה ממדים לממד אחד, וזאת באמצעות המכפלה הסקלרית של הווקטור העצמי המייצג את ה-PC הראשון בווקטור המשתנים המסבירים. למשל, נניח שווקטור המשתנים המסבירים של תצפית מסוימת i, לאחר שעברה תהליך סטנדרטיזציה, הוא: $x_i = (-0.43, 0.74, 0.52)$. ניתן לייצג את הווקטור הזה באמצעות משתנה מסביר אחד המתקבל כתוצאה של המכפלה:

$$u_1' \cdot x_i = [0.65 \quad 0.51 \quad 0.57] \cdot \begin{bmatrix} -0.43 \\ 0.74 \\ 0.52 \end{bmatrix} = 0.95$$

בבעיית החיזוי ניתן עכשיו לייצג את תצפית i באמצעות הערך היחיד x_i = 0.95 במקום שלושת המשתנים המסבירים המקוריים.

אם חוזרים על תהליך החישוב הזה עבור כל תצפית, מקבלים סדרת תצפיות עם משתנה מסביר אחד שאותו אומדים באמצעות מודל הרגרסיה.

RBF – Radial Basis Functions

RBF הוא אלגוריתם השייך למשפחת הרשתות העצביות (Neural Networks - רשתות נוירונים). רשת RBF היא רשת שמשתמשת בפונקציות בסיס רדיאליות (RBF – Radial Basis Functions) בתור פונקציות הפעלה (activation function). פונקציית בסיס רדיאלית $h(x)$ היא פונקציה שעבורה הערך של הפונקציה בנקודה מסוימת x תלוי במרחק של הנקודה מנקודת המרכז של הפונקציה c, כך שמתקיים $h(x-c) = h(||x-c||)$. (הסימון $||x-c||$ מסמן מרחק של הנקודה x מהמרכז c). כל פונקציה שמקיימת תנאי זה יכולה לשמש כפונקציית בסיס. הפונקציה הנפוצה ביותר, שבה השתמשנו גם במחקר זה, היא הפונקציה הגאוסיאנית (הנורמלית) (Powell, 1987) שבמקרה ה - k ממדי ניתנת על ידי הביטוי הבא:

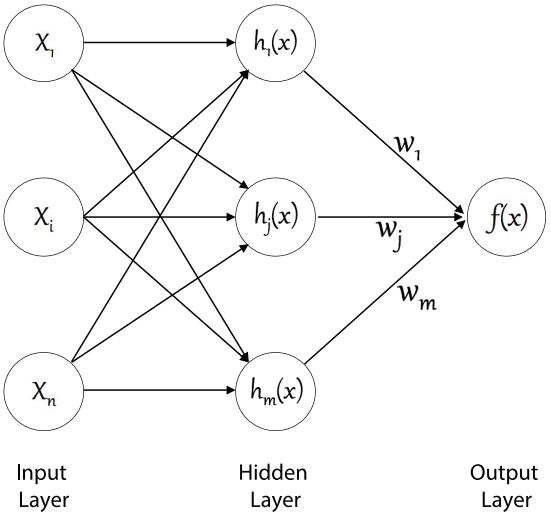
$$h_j(x) = \frac{1}{(2\pi\sigma_j^2)^{k/2}} exp\left\{-\frac{||x-\mu_j||^2}{2\sigma_j^2}\right\}$$

כאשר μ_j היא נקודת המרכז (התוחלת) של פונקציית הבסיס j

σ_j הוא ה"רוחב" שלה (סטיית התקן).

דוגמה של רשת RBF מופיעה באיור 3.

איור 3: רשת RBF עם m פונקציות בסיס



ברשת זו:

$x_1, x_2, \dots, x_n$ הם נתוני הקלט (התצפיות), כשכל תצפית מורכבת מ-K משתנים מסבירים.

$h_1(x), h_2(x), \dots, h_m(x)$ הם הפלטים של m פונקציות הבסיס שמתקבלים מתוך הפונקציה הגאוסיאנית שלעיל.

$w_1, w_{2,\dots,} w_m$ הם המשקלות של הרשת.

$f(x) = \sum_{j=1}^{m} w_j h_j(x)$ הוא הפלט של הרשת המתקבלת כקומבינציה לינארית של המשקלות והפלט של פונקציות הבסיס.

בדומה לרשתות עצביות, גם רשתות RBF מורכבות ממספר שכבות – שכבת קלט, שכבה נסתרת ושכבת פלט. וכמו ברשתות העצביות, גם רשתות RBF משמשות לחיזוי של אירועים על בסיס תצפיות מהעבר. אלא שלהבדיל מהרשתות העצביות המסורתיות שבהן מתבצע תהליך של למידה מונחית לצורך עדכון המשקלות של כל הענפים ברשת, שבו לוקחים חלק כל המשתנים המסבירים המקוריים, ברשתות RBF מתבצע תהליך מקדים של צמצום ממדי הבעיה מ-K המשתנים המקוריים ל-m משתנים חדשים, m<<K, כאשר m הוא מספר פונקציות הבסיס ברשת. תהליך זה מתבצע בשכבה הנסתרת של רשת ה-RBF שממפה את צפיפות התצפיות בקובץ הנתונים לאוסף של m פונקציות בסיס ומגדירה m משתנים מסבירים חדשים לכל תצפית, משתנה אחד עבור כל פונקציית בסיס. תהליך החיזוי מתבצע רק לאחר התהליך המקדים הזה ולוקחים בו חלק רק m המשתנים החדשים.

תהליך היישום של רשת RBF לבעיית חיזוי כולל שני שלבים, אמידה ורגרסיה.

בשלב הראשון מוצאים את הערכים האופטימליים של הפרמטרים של פונקציות הבסיס – נקודות המרכז (μ_j) והרוחב (σ_j), כך שתייצגנה בצורה הטובה ביותר את הפיזור של תצפיות הקלט במרחב הרב-ממדי של המשתנים המסבירים. במחקר שלנו אמדנו את הערכים האופטימליים של פונקציות הבסיס באמצעות אלגוריתם expected EM) maximization) (Dempster et al., 1977). בהינתן האומדים של הפרמטרים של פונקציות הבסיס, ניתן לחשב, לגבי כל תצפית, את הפלטים $h_j(x)$ של פונקציות הבסיס ואז להשתמש בערכים $h_j(x)$ כדי להחליף את המשתנים המסבירים המקוריים במשוואת הרגרסיה. בצורה זו אנחנו מקטינים את הממד של כל תצפית מ-K ממדים (מספר המשתנים המסבירים המקוריים) ל–m ממדים, כמספר פונקציות הבסיס.

בשלב השני של התהליך מריצים את מודל הרגרסיה הלינארית על תצפיות הקלט תוך שימוש בערכים של $h_j(x)$ שחושבו לעיל בתור המשתנים המסבירים החדשים (שמספרם עכשיו ירד באופן משמעותי מ–K משתנים ל–m משתנים) ומוצאים את מקדמי הרגרסיה w_j עבור כל אחד מהמשתנים.

נציין שתהליך מיפוי התצפיות שמתבצע בשכבה הנסתרת אינו אלא תהליך של הקבצה לאשכולות (clustering), כך שניתן להתייחס אל פונקציות הבסיס כאל אשכולות. אלא שבניגוד לגישות ההקבצה המסורתיות, כגון גישת K-Means, המשייכות כל תצפית לאשכול אחד בלבד, רשת RBF משייכת כל תצפית לכל אחת מפונקציות הבסיס על פי מרחק התצפית מנקודת המרכז של פונקציית הבסיס. יתרה מזו, ניתן גם לנרמל את פונקציות הבסיס כך שהפלט שלה, $h_j(x)$, ייצג את ההסתברות האפוסטריאורית שהתצפית שיכת לפונקציית הבסיס. עוד נציין שחלוקה לאשכולות היא אחת השיטות המקובלות לצמצום ממדים. במקרה הפשוט ביותר ממירים את K המשתנים המקוריים בסדרה של m משתני 0/1 (כאשר m הוא מספר האשכולות): 1 – אם התצפית שיכת לאשכול, 0 – אם אינה שייכת. ברור שבגישה זו יש הפסד רב של מידע שעשוי לעוות את תוצאות החיזוי. במקרה שלנו אנו משתמשים בהסתברויות האפוסטריאוריות כדי להגדיר את המשתנים החדשים, הסתברויות שלא זו בלבד שהן עשירות יותר במידע אלא יש להן גם בסיס תיאורטי.

בתור דוגמה מספרית, נניח בעיית חיזוי עם 3 משתנים מסבירים ועם פונקציית הפעלה רדיאלית תלת- ממדית, שבלי להגביל את הכלליות היא בעלת רוחב σ=1 עם נקודת מרכז בראשית (0,0,0)=μ. נבחר תצפית xi עם הערכים הבאים (לאחר סטנדרטיזציה) (0.1,0.1,0.1)=xi. ריבוע המרחק האוקלידי של תצפית זו מהמרכז הוא: $d^2 = (0.1^2 + 0.1^2 + 0.1^2) = 0.03$. אם נציב בנוסחת ההתפלגות הגאוסיאנית הנ"ל עם שונות 1, נקבל:

$$h(x_i) = \frac{1}{(2\pi)^{3/2}} exp\left\{-\frac{0.03}{2}\right\} = 0.063$$

עכשיו ניתן לייצג את התצפית התלת-ממדית באמצעות המשתנה היחיד X_i= 0.063 ובכך להקטין את הבעיה מ-3 ממדים לממד אחד.

את תהליך החישוב הזה יש להפעיל על כל התצפיות בקובץ הקלט ואז מקבלים סדרה חד-ממדית של תצפיות, והיא זו שעוברת לשלב השני של התהליך (מודל הרגרסיה).

ברשת שמכילה מספר פונקציות בסיס, למשל 3, יש צורך לחשב, עבור כל תצפית, את הערך $h_j(x)$ עבור כל אחת מפונקציית בסיס ולהמיר את K המשתנים המסבירים המקוריים ב–3 הערכים של $h_j(x)$.

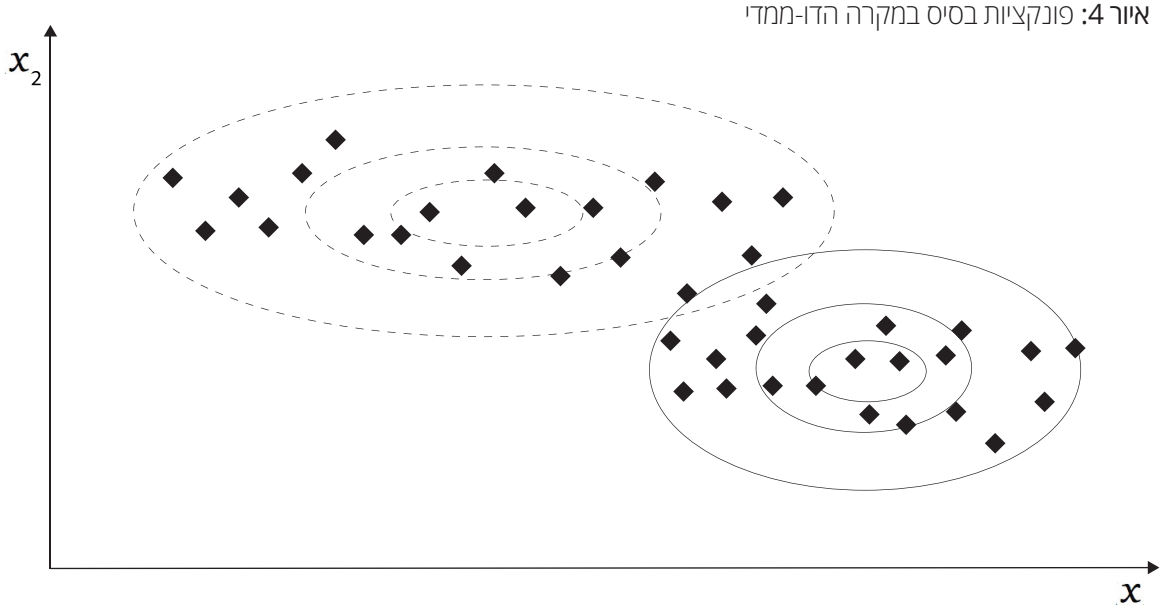
איור 4 מדגים את גישת RBF במרחב הדו–ממדי.

הנקודות באיור 4 מייצגות את התצפיות שמיפינו באמצעות שתי פונקציות בסיס רדיאליות דו-ממדיות המתוארות באיור באמצעות קווי הגובה של פונקציות הצפיפות (מעגליות או אליפטיות) המרכיבות אותן. פונקציית הצפיפות לאורך כל קו גובה קבועה והיא הגבוהה ביותר במרכז. ההסתברות האפוסטריאורית של כל נקודה נקבעת על סמך מיקומה במערכת קווי הגובה.

יישום גישת RBF לצמצום ממדי הבעיה ופתרון בעיית החיזוי מורכבים מהשלבים הבאים:

- אתחול – בחרו את מספר פונקציות הבסיס m ואת המרכזים שלהם, לדוגמה, באמצעות אלגוריתם K ממוצעים (K-Means algorithm) או בצורה אקראית.
- אימון - הריצו את אלגוריתם EM, בדרך כלל תוך שימוש בכל התצפיות בקובץ האימון, כדי למצוא את הערכים האופטימליים של המרכז והרוחב של כל פונקציות הבסיס.
- חישוב הסתברויות – חשבו את ההסתברויות האפוסטריאוריות $h_j(x)$ עבור כל התצפיות.
- רגרסיה – השתמשו בהסתברויות האפוסטריאוריות שחושבו לעיל בתור המשתנים המסבירים, הריצו את מודל הרגרסיה ומצאו את המשקלות w_j.
- תיקוף וחיזוי – חשבו את ההסתברויות האפוסטריאוריות עבור כל התצפיות בקובץ התיקוף והציבו אותן בנוסחת הרגרסיה כדי לחזות את הערך של משתנה התגובה.

איור 4: פונקציות בסיס במקרה הדו-ממדי



נציין שאמידת הפרמטרים של פונקציות הבסיס בשלב האימון, באמצעות אלגוריתם EM, היא בעיה של הקבצה לאשכולות, שהיא בעיית למידה בלתי מונחית. לעומתה, אמידת המשקלות בשלב הרגרסיה היא בעיית למידה מונחית.

בדומה לגישת PCA, גם בגישת RBF יש צורך לעשות סטנדרטיזציה של משתני המקור כדי למנוע עיוות של התוצאות.

השוואת השיטות

כזכור, מטרת מחקר זה היא להשוות את השיטות השונות שנדונו לעיל במטרה למצוא את השיטה הטובה ביותר לבחירת משתנים מסבירים במודלים של חיזוי. לצורך ההשוואה השתמשנו בשלושה בסיסי נתונים מתחום השיווק שסופקו על ידי DMEF (Direct Marketing Educational Foundation) (https://www.marketingedge.org/marketing-programs/data-set-library):

קובץ Non-Profit מתייחס למבצע שיווקי שמטרתו לגייס תרומות לארגון צדקה עם משתנה תלוי בינרי 0/1: 1 – הצרכן הגיב למבצע ותרם לארגון, 0 – הצרכן לא הגיב.

קובץ Specialty מתייחס למבצע שיווק של קטלוג מוצרים עם משתנה תלוי רציף - שיעור התגובה (מספר ההזמנות חלקי מספר הקטלוגים שנשלחו לצרכן).

קובץ Gift שמתייחס לקטלוג של מתנות עם משתנה תלוי בדיד ובר מנייה – מספר ההזמנות מהקטלוג.

מספר מאפיינים של בסיסי הנתונים האלה מופיעים בטבלה 2.

טבלה 2 מפרטת עבור כל קובץ את שם הקובץ, מספר התצפיות, מספר התגובות במבצע השיווק, מספר המשתנים המקוריים, מספר המשתנים לאחר הפעלת טרנספורמציות על המשתנים המסבירים המקוריים וסוג משתנה התגובה.

כאמור לעיל, הבסיס של מודל החיזוי במחקר זה היה מודל הרגרסיה הלינארית. לצורך הבדיקה חילקנו כל קובץ באופן אקראי לקובץ אימון ולקובץ תיקוף כשכל קובץ מכיל מחצית מהתצפיות. את מודל החיזוי בנינו על סמך קובץ האימון ובדקנו את התוצאות על קובץ התיקוף. עבור כל קובץ בנינו 4 מודלים של חיזוי, מודל אחד לכל אחת מהשיטות לבחירת מאפיינים שפורטו לעיל. כדי להביא את כל השיטות לבחירת משתנים לאותו בסיס השוואה, ולהשוות "תפוחים לתפוחים", מצאנו את הקונפיגורציה האופטימלית עבור כל גישה ואז השווינו את המודלים המיטביים זה לזה. תהליך האופטימיזציה כלל חיפוש על מרחב הפרמטרים הרלבנטיים לכל גישה, בטווחים שונים. כמו כן, בדקנו גם מספר פונקציות של טיב התאמה עבור כל גישה, הכוללות את מקדם הקביעה המותאם ($\bar{R}^2$ - adjusted r-square), ופונקציות טיב התאמה שנגזרות מקריטריון ה-AIC (Akaiki Information Criterion). נציין שאת מודל SWR כיילנו באמצעות שיטת FDR באמצעות האלגוריתם שתואר לעיל. עוד נציין שבשלב זה התעלמנו מטרנספורמציות לא לינאריות של משתנים היכולים לתת עדיפות לגישה זו או אחרת והפעלנו את הגישות השונות לבחירת משתנים רק על המשתנים המקוריים. את תהליך האופטימיזציה הזה הפעלנו לגבי כל גישה בנפרד על כל אחד מבסיסי

טבלה 2: המאפיינים של בסיסי הנתונים המשתתפים בתהליך ההשוואה

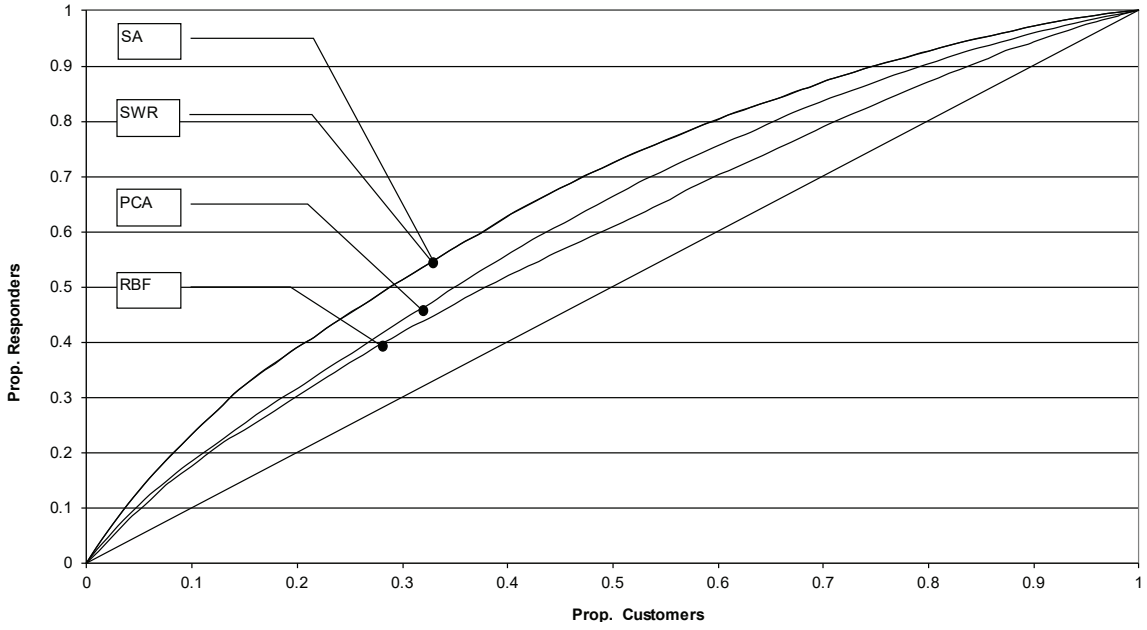
Name	#Observ.	# of Resp.	#Initial Var.	# Total pred.	Response Variable
Non-Profit	99,200	27,208	77	307	Binary
Specialty	106,284	5,758	287	350	Continuous
Gift	101,532	9,707	99	104	Counter

הנתונים הנ"ל, כדי למצוא את המשתנים המסבירים הטובים ביותר לכל מודל, היינו את אלה שמשיאים את פונקציית טיב ההתאמה. בשלב הסופי בנינו 4 מודלים של רגרסיה, מודל אחד לכל אחת מהגישות לבחירת משתנים, עבור כל אחד משלושת בסיסי הנתונים שהשתתפו במחקר.

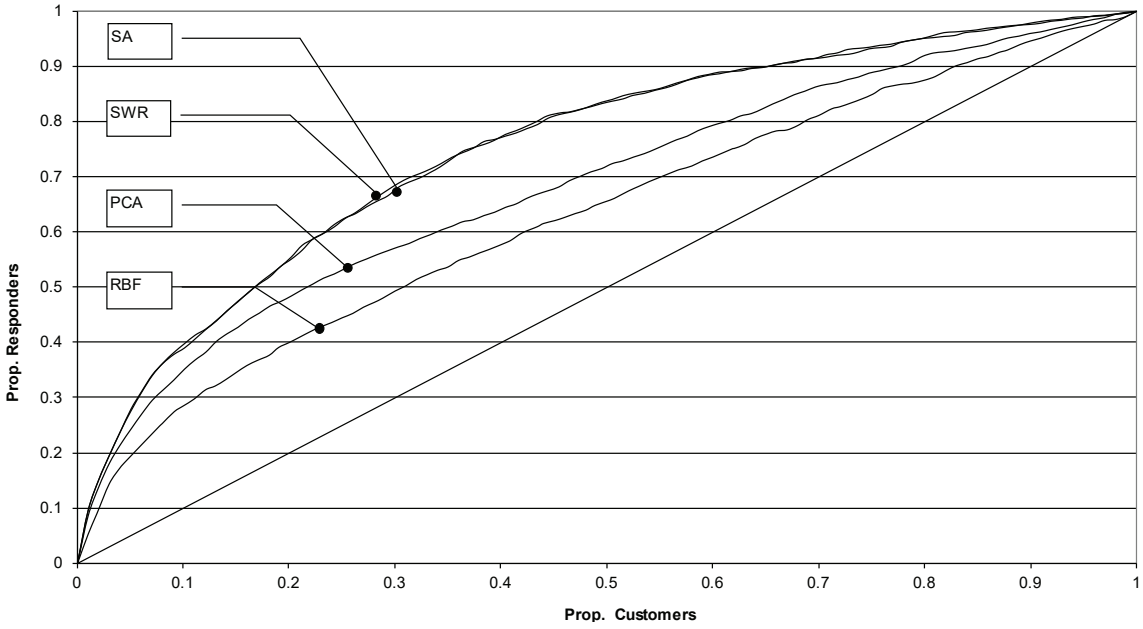
לצורך ההשוואה בין המודלים השתמשנו במספר מדדים מקובלים להערכת מודל חיזוי, ובהם דיאגרמת רווחים (gains chart), מספר המשתנים המסבירים שנכנסו למודל הסופי, R^2 עבור קובץ האימון, R^2 עבור קובץ התיקוף, היחס של R^2 בין קובץ האימון לקובץ התיקוף, מדד ג'יני (Gini coefficient), וה-lift המקסימלי (M-L, Maximum Lift). כדי למנוע מצב של התאמת יתר דרשנו מכל מודל שהיחס של מקדם הקביעה במדגם האימון ובמדגם התיקוף יהיה קרוב ל-1. דרישה זו חייבה אותנו להרחיב את מרחב החיפוש שיבטיח לא רק שנמצא את מערכת הפרמטרים הטובים ביותר לכל גישה של בחירת משתנים, אלא גם שנקבל מודל המקיים את האילוץ על מקדם הקביעה. דיאגרמת הרווחים היא המדד המפורט ביותר להערכת טיב החיזוי. ציר ה-x של דיאגרמת הרווחים מפרט את פרופורציות התצפיות (במקרה שלנו – צרכנים שהשתתפו במבצע השיווק) כשהתצפיות מסודרות בסדר יורד של התחזית של המשתנה התלוי (למשל הסתברות התגובה בקובץ ה-non-profit), וציר ה-y את פרופורציה המגיבים (למשל, פרופורציה הקונים בקובץ ה– non-profit). בדרך כלל נהוג להציג את נתוני החיזוי ברמה של עשירונים כשהקו האלכסוני המחבר את הנקודה (0,0) עם הנקודה (1,1) הוא מודל האפס (null model) שמשמש כנקודת הייחוס להערכת טיב החיזוי. מדד M-L מייצג את ההבדל המקסימלי בין מודל האפס לעקום התגובה על פי מודל החיזוי ומדד ג'יני את השטח הכלוא בין עקום התגובה ובין מודל האפס.

לצורך נוחות ההשוואה אנחנו מציגים את דיאגרמת הרווחים עבור ארבעת המודלים (SWR, SA, PCA, RBF) אחד על גבי השני עבור כל אחד מקובצי הנתונים שלקחו חלק במחקר (איורים 5-7). כזכור, כל דיאגרמת רווחים מתייחסת לקונפיגורציה האופטימלית של המודל המתאים. טבלה 3 מרכזת את את כל מדדי הביצוע עבור המודלים השונים. נציין שכל דיאגרמות הרווחים וכן מדד M-L ומדד ג'יני מתייחסים רק לקובץ התיקוף.

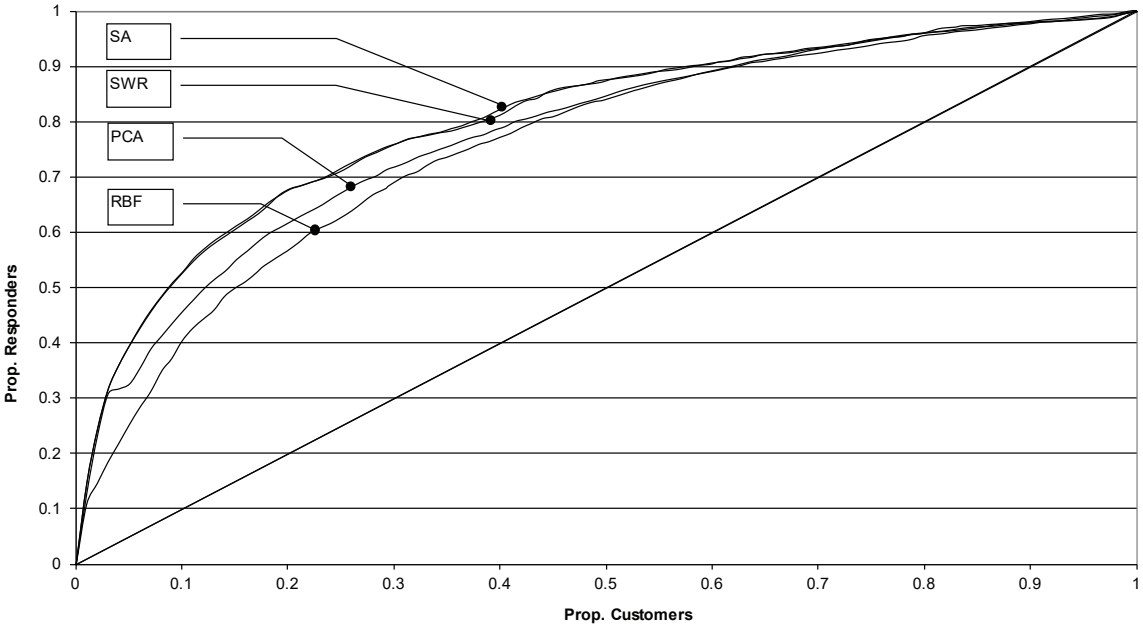
איור 5: דיאגרמות רווחים עבור קובץ Non-Profit



איור 6: דיאגרמות רווחים עבור קובץ Specialty



איור 7: דיאגרמות רווחים עבור קובץ Gift



טבלה 3: השוואה בין המודלים

File	Parameter	RBF	PCA	SA	SWR
Non-Profit	Number of predictors	40	40	25	20
	R^2 Training	0.0415	0.0645	0.1245	0.1250
	R^2 Validation	0.0380	0.0639	0.1196	0.1201
	R^2 – Ratio	0.9148	0.9905	0.9600	0.9940
	Gini	0.1695	0.2301	0.3155	0.3154
	M-L	0.1262	0.1671	0.2362	0.2285
Specialty	Number of predictors	50	50	28	22
	R^2 Training	0.0290	0.0455	0.0710	0.0722
	R^2 Validation	0.0254	0.0443	0.0650	0.0650
	R^2 – Ratio	0.8747	0.9729	0.9157	0.9000
	Gini	0.2666	0.3739	0.5002	0.5018
	M-L	0.2013	0.2853	0.3816	0.3837
Gift	Number of predictors	40	50	34	31
	R^2 Training	0.1602	0.2305	0.3426	0.3412
	R^2 Validation	0.1609	0.2214	0.3383	0.3388
	R^2 – Ratio	1.0043	0.9605	0.9874	0.9542
	Gini	0.5194	0.5584	0.6114	0.6093
	M-L	0.3911	0.4250	0.4797	0.4797

מהתבוננות בדיאגרמות הרווחים עולה המסקנה המפתיעה שעבור כל שלושת קובצי הנתונים שלקחו חלק במחקר, גישת SWR נותנת תוצאות כמעט זהות על קובץ התיקוף בהשוואה לגישת SA. מדוע מפתיעה? כיוון שעל פניה גישת SA נחשבת חזקה יותר בשל יכולתה לדלג על נקודות מקסימום מקומי ולהתכנס לנקודת המקסימום הגלובלית, וזאת בניגוד לגישת SWR שהיא גישה מיופית (myopic) שבכל איטרציה מתקדמת רק צעד אחד קדימה עם סיכוי גבוה להיתקע בנקודת מינימום מקומי. גם כשגישת SA מניבה מדדים טובים יותר מגישת SWR (למשל מדד ג'יני גבוה יותר עבור קובץ non-profit וקובץ gift), ההבדלים הם מזעריים ולא משמעותיים. יתרה מזו, בחינה של המשתנים המסבירים שנכנסו למודל מראה שבשתי הגישות יש חפיפה גדולה בין המשתנים שנבחרו למודל הסופי וגם מספר המשתנים שנכנסו למודל הוא דומה. העובדה שהתקבל פתרון דומה בשתי גישות השונות זו מזו לחלוטין יכולה להעיד על כך שהפתרון שקיבלנו קרוב מאד לאופטימום הגלובלי. אפילו אם הפתרונות אינם זהים ממש, הם נמצאים על אותה רמה (platuea) ועל כן מניבים מדדי ביצוע דומים.

מה מסביר את התופעה שבה גישת SWR, שעל פניה היא נחותה מגישת SA בשל היותה גישה מיופית, נותנת לבעיית בחירת המשתנים המסבירים למודל החיזוי פתרון כמעט זהה לפתרון של גישת SA המתוחכמת יותר? ככל הנראה ההסבר לכך הוא שבמחקר זה התמקדנו רק בתחום השיווק ובדקנו רק קובצי נתונים של מבצעי שיווק. ההנחה היא שצרכנים בעולם השיווק הם עקביים ומקבלים החלטות קנייה בצורה עקבית ורציונלית. התנהגות רציונלית זו, כשהיא משתקפת בקובצי הנתונים, גורמת לכך שקובצי הנתונים בשיווק מתנהגים היטב ללא קשרים לא לינאריים ו/או קשרים מורכבים אחרים בין המשתנים המסבירים למשתנה התלוי (כגון קשרים מסוג XOR). פועל יוצא של תכונה זו המאפיינת את תחום השיווק הוא שגם גישה מיופית כמו גישת SWR מסוגלת למצוא פתרון לבעיית בחירת המשתנים המסבירים למודל חיזוי שאיננה נחותה מהפתרונות המתקבלים בגישות מתוחכמות יותר.

כדי לאשש מסקנה זו בדקנו את שתי הגישות גם על שני קבצים סינתטיים שיצרנו באמצעות סימולציה, ואשר הכילו קשרים

מורכבים בין המשתנים המסבירים למשתנה התלוי. כל אחד מקבצים אלה הכיל 100,000 תצפיות אך באחד מהם היו 100 משתנים מסבירים ובשני 200. כדי לבדוק את רגישות התהליך הֵרצנו את גישות SWR ו-SA על מספר קונפיגורציות של פרמטרים. התוצאות שהתקבלו היו חד-משמעיות לטובת גישת SA, ואפילו מפתיעות. הסתבר שגישת SWR נכשלה לחלוטין על קבצים אלה. לא זו בלבד שלא הצליחה להפריד בין הצרכנים ה"טובים" לצרכנים ה"לא טובים" אלא גם הניבה עקום תגובה שהתנודד מעל ומתחת למודל האפס עם דיאגרמת רווחים שנראתה יותר מכול כמו רעשים מקריים. לעומתה, גישת SA הצליחה לדלג על הפתרונות הלוקליים ולתת פתרון שהצליח לאפיין את הצרכנים הטובים עם עקום תגובה בדיאגרמת הרווחים שנמצא מעל פתרון האפס. (תיאור מפורט של תהליך הסימולציה וגישת SA ניתן למצוא אצל Me'iri and Zahavi (2006.

אשר לגישות PCA ו-RBF, התבוננות בדיאגרמות הרווחים מעידה שהן נחותות ממש יחסית לגישות SWR ו-SA שכן עבור כל קובצי הנתונים שהשתתפו במחקר, דיאגרמות הרווחים שלהם נמצאים מתחת לדיאגרמות הרווחים של גישת SWR ושל גישת SA. אנו מייחסים תופעה זו לעובדה שגישות PCA ו-RBF הן גישות לצמצום ממדים המתחשבות רק במשתנים המסבירים ללא כל התייחסות למשתנה התלוי, ולכן הן נותנות פתרונות חלשים יותר. בנוסף, שתי הגישות האלה מייצגות את המשתנים המסבירים באמצעות מספר "משתני על" שמתקבלים באמצעות טרנספורמציות של המשתנים המקוריים, מה שמקשה על האינטרפרטציה של תוצאות המודל.

אין ספק שלמסקנות אלה משמעות מעשית חשובה שכן משתמע מהן שבעולם השיווק ניתן להשתמש בגישות SWR כדי לבנות מודלים של חיזוי במקום במודלים מתוחכמים יותר כגון SA. לא רק שגישות SWR מוכרות ביותר בעולם העסקי (וגם נלמדות בבתי ספר לניהול...), אלא שהן נמצאות גם בכל התוכנות הסטטיסטיות המובילות כגון SAS, SPSS, R ואחרות. זאת ועוד, מחקרים נוספים שערכנו מעידים שגישת SWR היא גמישה מאוד ונותנת פתרון סביר עבור טווח רחב של פרמטרים, עובדה המקלה גם על המשתמש העסקי, שאינו בקיא בסטטיסטיקה, לבנות מודלים של חיזוי בכך שאינה מחייבת חיפוש על מרחב הפרמטרים כדי למצוא את הפתרון המיטבי.

ראוי לציין שבמחקר זה התמקדנו רק בשלוש שיטות מובילות לבניית מודלים רב-ממדיים לצורך חיזוי, מתוך אוסף רחב יותר של שיטות שפותחו לאחרונה הכוללות עצי החלטה, random forest, רשתות נוירונים מסוגים שונים, SVM, גישות חדשות של למידה עמוקה ואחרות. אבל לאור תוצאות מחקרנו נראה ששימוש בשיטות אלה לצורכי בעיות שיווק הוא "מינון יתר" מופרז, בשל הממצא המרכזי שלנו שהאלגוריתם הפשוט יחסית של רגרסיה בצעדים נותן פתרון לבעיית בחירת המשתנים המסבירים למודל רגרסיה שאינו נחות מהפתרונות המתקבלים באמצעות שיטות מתוחכמות אחרות, שמן הסתם גם קשות יותר ליישום. אין ספק שיש מקום לבדוק שיטות אלה במחקר עתידי, בוודאי בתחומים אחרים כגון מימון, IOT, גילוי הונאות, סייבר ועוד. אולם על פי תוצאות המחקר שלנו נראה שככל שמדובר בבעיות שיווק, הגישה המנצחת היא רגרסיה... בצעדים (SWR).

סיכום

במחקר זה בדקנו מספר גישות לאוטומציה של תהליך בחירת המשתנים עבור מודלים של רגרסיה בעולם נתוני העתק שמכילים מאות, ואף יותר, משתנים מסבירים. במחקר התמקדנו בשלוש גישות לאוטומציה של התהליך – גישות סטטיסטיות, גישות סטוכסטיות וגישות של צמצום ממדים. בכל גישה בחרנו מודל מייצג אחד או שניים, ובסך הכול 4 מודלים: רגרסיה בצעדים (SWR), Simulated Annealing (SA), Principal Component Analysis (PCA) ו-Radial Basis Functions (RBF). עבור כל מודל מצאנו את הקונפיגורציה האופטימלית והשווינו את המודלים המיטביים שהתקבלו זה לזה על מספר קובצי נתונים מתחום השיווק. התוצאות שהתקבלו מעידות שלפחות בתחום השיווק גישת SWR נותנת תוצאות שאינן נחותות מגישות שעל פניהן נראות עדיפות ומתוחכמות יותר, כגון גישות SA. למסקנה זו משמעות מעשית חשובה שכן משתמע ממנה שכדי לבנות מודלים רב-ממדיים של חיזוי בעולם השיווק ניתן להסתפק בגישת SWR הטובה והמוכרת הזמינה בכל התוכנות הסטטיסטיות המובילות.

אין ספק שאוטומציה של בחירת משתנים למודל חיזוי היא חלק הכרחי בתהליך ה"דמוקרטיזציה" של מודל החיזוי אם רוצים להנגיש את התהליך גם למשתמשים עסקיים

ולמשתמשים חסרי רקע מעשי או תיאורטי בסטטיסטיקה ובכריית מידע. אולם בחירת המשתנים למודל היא רק חלק מהתהליך. מרכיבים נוספים הם העיבוד המקדים לשם הגדרה וטיוב המשתנים המסבירים, טיפול בטרנספורמציות ובאינטראקציות בין משתנים, בדיקה ותיקוף של המודלים ופתרון בעיות של התאמות יתר, יישום המודלים על תצפיות חדשות (תהליך ציינון - scoring) ועוד. כמו כן, קיימים הבדלים בין מודלים להסבר של תופעות ובין מודלים של חיזוי אנליטי שיש להם השפעה על אופן בניית מודלים לחיזוי ועל יישומם. תקצר היריעה מלדון בסוגיות אלה במאמר זה. דיון נרחב יותר בתהליך החיזוי האנליטי ובהבדלים בין מודל הסבר למודל חיזוי מופיע במאמר על חיזוי אנליטי (יעקב זהבי, 2017).

במאמר זה התמקדנו בתחום השיווק והגענו למסקנה שגישת SWR נותנת מענה טוב לבעיית בחירת המשתנים המסבירים למודל רגרסיה. עם זאת, בעיות חיזוי הן בעיות מגוונות ביותר וקרוב לוודאי שבתחומים אחרים משיווק, שיטת SWR לא תיתן בהכרח את הפתרון המיטבי לסוגיית בחירת המשתנים המסבירים למודל. יתרה מזו, לא מן הנמנע שבתחומים אחרים יהיה צורך לבחון מספר גישות, מתוך מגוון הגישות לבעיות חיזוי שחלקן גם הוזכרו לעיל, כדי למצוא את המודל שיספק את התוצאה המיטבית. זאת כיוון שלא תמיד ניתן לדעת מראש מהו המודל שיפיק את התוצאה הטובה ביותר. עובדה זו, נוסף על כל הסוגיות שפורטו לעיל, תקשה עוד יותר על תהליך ה"דמוקרטיזציה" של מודלים לחיזוי. אבל על אף הקשיים, הרכבת כבר יצאה מהתחנה והציפייה היא שתוך מספר שנים ניתן יהיה להפוך את תהליך החיזוי האנליטי לנגיש, זמין ונוח לקהל משתמשים רחב.

פרופ' יעקב זהבי jacobz@tauex.tau.ac.il

רשימת מקורות

Benjamini ,Y., & Yekutieli, D. (2001). The Control of the False Discovery Rate in Multiple Testing under Dependency. Ph.D. thesis OR department Tel Aviv University.

Benjamini, Y., & Hochberg, Y. (1995). Controlling the False Discovery Rate: a Practical and Powerful Approach to Multiple Testing. J. R. Statist. Soc. B (1995) 57 (1), 289-300.

Dempster, A. P., Laird N. M., & Rubin D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society, B 39 (1), 1-38.

Dunteman, G. H. (1994). Principal Components Analysis in Factor Analysis and Related Techniques, Part III: (Editor Lewis-Beck M. S.), SAGE Publication Toppan Publishing.

Efoymson, M.A. (1960). Multiple Regression Analysis in Mathematical Method for Digital computers, Wiley, New York, 191-203.

Geman, S., Bienenstock, E., & R. Doursat. (1992). Neural Networks and Bias/Variance Dilemma. Neural Computation 4 (1), 1-58.

Goldberg, D. E. (1989). Genetic Algorithms in Search Optimization and Machine Learning. Addison-Wesley Publishing Company Inc.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). the Elements of Statistical Learning, Springer-Verlag, NY.

Hochberg, Y., & Tamhane, A. C. (1987). Multiple Comparison Procedures. John Wiley & Sons.

Levin, N., & Zahavi, J. (1977). Applying Neural Computing to Target Marketing, Journal of Direct Marketing, Vol. 11, 5-22.

Me'iri, R., & Zahavi, J. (2006). Using Simulated Annealing to Optimize the Feature Selection Problem in Marketing Applications, the European Journal of Operations Research, 171, 842-858.

Miller, A. J. (2002). Subset Selection in Regression, Chapman and Hall., London.

Powell, M. J. D. (1987). Radial Basis Functions for Multivariable Interpolation: a Review. in Algorithms for Approximation (Eds. Mason J. C. & Cox M.G.).

Tukey, J.W. (1977). "Some Thoughts on Clinical Trials, especially Problems of Multiplicity", Science, 198, 679-684.

Van Laarhoven, P.J.J., & Aarts, E.H.L. (1987). Simulated Annealing Theory and Application, Kluwer, Boston.

זהבי, י'. חיזוי אנליטי (Predictive Analytics) – הלכה למעשה (2017), חידושים בניהול, הפקולטה לניהול ע"ש קולר, אוניברסיטת תל אביב, 1, 55-69.