הפקולטה לניהול
ע"ש קולר
אוניברסיטת תל אביב

מאמר זה פורסם בכתב העת "חידושים בניהול", גיליון מס' 13, ספטמבר 2023,
היוצא לאור על ידי הפקולטה לניהול על שם קולר באוניברסיטת תל אביב.

יישומים של ניתוח הישרדות בתחומי הניהול ומדעי החברה

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. החל את הקריירה המקצועית שלו בתחום 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 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) המאורגנת על ידי ACM (American Computation Machinery). מספר מאמרים שלו בתחום זה זכו בפרסים על מצוינות אקדמית.

תקציר

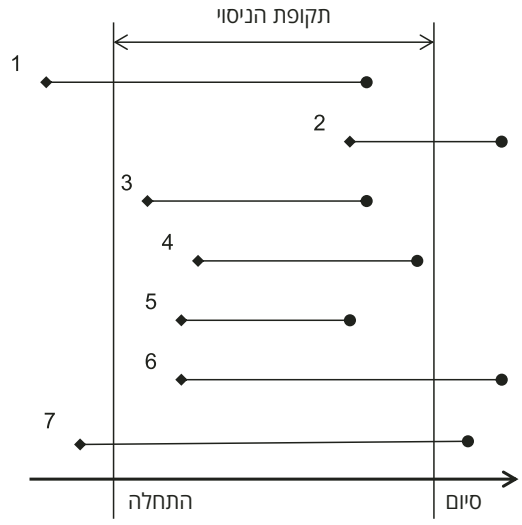
ניתוח הישרדות (Survival Analysis) הוא תחום בסטטיסטיקה העוסק בחיזוי סיכויי ההישרדות של אירועים הקורים לאורך זמן במהלך תקופת ניסוי (Experimental Period). המקור של ניתוח ההישרדות הוא מתחום הרפואה, שבו פונקציית המטרה היא לאמוד את הסיכוי שזמן ההישרדות עד לתמותה או להחלמה ממחלה מסוימת, בתקופת הניסוי, גדול מ-t יחידות זמן. עם זאת, אירועים תלויי זמן שכיחים מאוד גם בתחומים אחרים – ניהול, מדעי החברה, הנדסה ועוד. לכן לניתוח הישרדות יש שימושים פוטנציאליים רבים גם בתחומים אלה. אלא שהדיון בספרות ביישומים של ניתוח הישרדות בתחומים שהם לא רפואה הוא דליל למדי. נוסיף לכך את העובדה שהמודלים של ניתוח הישרדות הם בין המודלים היותר מורכבים בסטטיסטיקה ולכן הם פחות מוכרים בתעשייה. כתוצאה מכך, תהליך בניית המודלים, גם באירועים תלויי זמן, מתבסס לרוב על מודלים "רגילים" של רגרסיה במקום מודלים של ניתוח הישרדות, מה שעשוי להביא למודלים מוטים ולהחלטות שגויות. מטרת המאמר הזה היא להציף את הפוטנציאל שיש למודלים של ניתוח הישרדות עבור אירועים תלויי זמן גם בתחומי הניהול ומדעי החברה, לעורר את המודעות לנושא הזה, ולהנגיש אותו לתעשייה. אנו גם מציגים פתרון שמשלב מודלים לניתוח הישרדות בתהליכים של חיזוי אנליטי המאפשר לחזות אירועים גם מעבר לתקופת הניסוי, ומדגימים אותו באמצעות בעיה מעשית בתחום הרכב.

1. הקדמה

ניתוח הישרדות, Survival Analysis, ובקיצור SA, הוא תחום בסטטיסטיקה העוסק בחיזוי סיכויי ההישרדות של אירועים הקורים לאורך זמן (Kleinback et al., 2007, Miller, 1997). המקור של SA הוא בתחום הבריאות והרפואה, שבו האירועים בעלי העניין הם אירועים רפואיים כגון תמותה, החלמה, אשפוז חוזר ועוד. לאחרונה נכנסו הכלים של SA לתחומים רבים נוספים.

בדרך כלל תהליך SA מתייחס לתקופת זמן מסוימת הנקראת תקופת הניתוח (Analysis Period) או תקופת הניסוי (Experimental Period). לדוגמה, במחקרים רפואיים תקופה טיפוסית במעקב אחרי חולי סרטן היא חמש שנים, כשהמטרה היא לאמוד את סיכויי ההישרדות מהמחלה לאחר מתן טיפול רפואי מסוים, כעבור שנה, שנתיים, שלוש שנים וכו' מיום תחילת הניסוי. מה שמאפיין סוג כזה של נתונים הוא תופעת הקיטום (Censoring) שנורמת להפסד אינפורמציה. למשל, לגבי חולים שחלו לפני תחילת תקופת הניסוי, אנו מאבדים את האינפורמציה על מה שקרה עד שהחולה הצטרף לקבוצת הניסוי. ובאשר לחולים ששרדו את מלוא תקופת הניסוי, אנו מאבדים את האינפורמציה על מה שקרה לאחר תום תקופת הניסוי. תופעת הקיטום מתרחשת כשהאירוע שבו מדובר לא מתממש עד לסוף תקופת הניסוי, או שכשהאירוע החל לפני תחילת תקופת הניסוי והתממש רק במהלך תקופת הניסוי (או אחריו).

הסוג הזה של קיטום שבו מגיעים לסוף תקופת הניסוי מבלי לדעת איך ואם האירוע המדובר הסתיים, הוא קיטום מימין (Right censoring), כלומר התופעה המדוברת ממשיכה לאחר תום תקופת הניסוי (כלומר מימין). הסוג השני הוא קיטום משמאל (Left censoring) המתייחס לתצפיות שנכנסו לתקופת הניסוי לפני התחלת תקופת הניסוי (כלומר משמאל), כגון אדם שחלה לפני תקופת הניסוי. תופעת הקיטום גורמת לאובדן אינפורמציה, שכן בשני המקרים של קיטום מימין וקיטום משמאל אנחנו לא יודעים במדויק את משך הזמן שקורה האירוע שבו מדובר (במקרה של קיטום מימין מפני שאנחנו לא יודעים מתי האירוע הסתיים, ובמקרה של קיטום משמאל מפני שאנחנו לא יודעים מתי האירוע התחיל). יש גם תצפיות עם קיטום הן מימין והן משמאל – כלומר האירוע שבו מדובר התחיל במועד לא ידוע, עוד לפני שהתצפית נכנסה לתקופת הניסוי, וגם הסתיים במועד לא ידוע, רק לאחר תום תקופת הניסוי. מצב של קיטום מרווחי (Interval censoring) מתייחס לאירועים שקרו בתוך תקופת הניתוח, אבל לא ידוע לנו בדיוק מתי הם קרו. לדוגמה, אדם שחלה וגם החלים בתוך תקופת הניסוי. דוגמאות של תצפיות קטומות מופיעות באיור 1.



איור 1: תופעת הקיטום

תצפיות 1 ו-7 הן תצפיות עם קיטום משמאל מכיוון שהן מתחילות לפני תקופת הניסוי. תצפיות 2, 6 ו-7 הן תצפיות עם קיטום מימין מכיוון שהן מסתיימות לאחר תום תקופת הניסוי. תצפיות 3, 4 ו-5 מייצגות קיטום מרווחי.

שיטות האמידה והחיזוי המקובלות, כגון שיטות רגרסיה, וכל הסטטיסטיים ומבחני ההשערות התומכים בתהליך הרגרסיה, אינם תופסים במקרה של נתונים קטומים מפני שהדרישה היא שכל התצפיות, כולל אלה הקטומות, ישתתפו בתהליך הסטטיסטי, אחרת יש כאן הפסד אינפורמציה. אולם התייחסות לנתוני זמן קטומים כאל זמני ההישרדות האמיתיים של התצפיות הקטומות יגרמו להטיות בזמני ההישרדות וההסתברויות שלהם. ולכן נדרשות גישות סטטיסטיות אחרות לטיפול בנתונים כאלה. כאן נכנס לתמונה נושא ניתוח ההישרדות שמציע מודלים סטטיסטיים שלוקחים בחשבון את תופעות הקיטום.

נבחין בין שני סוגים של ניתוח הישרדות על פי סקלת המדידה של המשתנה התלוי – נומינלי (בדיד) ונומרי (רציף). במקרה הנומינלי, המשתנה התלוי הוא בדיד ובר מנייה, כגון מספר

מחלימים, מספר פטירות, מספר פצועים וכדומה. המשתנה הנומינלי המוביל הוא משתנה התמותה, שעומד ביסוד של שורת מחקרים בתחומים כמו אקטואריה, דמוגרפיה, מחקרים אפידמיולוגיים, בריאות הציבור (Public health), גאוגרפיה ועוד. אבל ניתוח ההישרדות ישים גם לשורה רחבה של משתנים נומינליים מתחומים שונים במדעי החברה והניהול, כגון מספר הזוגות שעדיין מתמידים בנישואים בתום תקופת הניסוי, מספר הזוגות שהתגרשו, מספר התינוקות שנולדו, גובה ההגירה החיובית והשלילית ברמה מקומית או ארצית, ועוד שימושים רבים. פונקציית המטרה במשתנים מסוג זה היא לאמוד את מספר האירועים בסוף תקופת הניסוי או אפילו בשלבים שונים של תקופת הניסוי.

במקרה הנומרי, המשתנה התלוי המוביל הוא הזמן, שהוא משתנה רציף והמטרה היא לאמוד את פונקציית ההתפלגות של משך הזמן עד לקרות האירוע. בהינתן פונקציית ההתפלגות אפשר לגזור ממנה פרמטרים שונים, כגון תוחלת (Expected value) משך הזמן, חציון (Median) משך הזמן, ועוד. למשל, באירועים רפואיים נרצה לאמוד את תוחלת משך הזמן עד להחלמה ממחלה מסוימת, תוחלת משך הזמן עד לתמותה, תוחלת משך הזמן עד לחזרה לאשפוז חוזר, ועוד. ביישומים אחרים נרצה לאמוד את תוחלת גיל הנישואין, תוחלת משך זמן הנישואין, תוחלת משך הזמן בין לידות עוקבות של אותה אישה, תוחלת משך הזמן שאדם גר באותו העיר או מתמיד באותה עבודה, ועוד.

במאמר זה אנו נשים את הדגש על המקרה שבו המשתנה התלוי הוא הזמן עד שקורה האירוע שבו מדובר, שהוא יותר אופייני לבעיות בתחום הניהול ומדעי החברה. במקרים רבים משך הזמן עד שקורה האירוע תלוי במשך הזמן שעבר מאז שקרה האירוע הקודם. למשל, משך הזמן עד לקניית מכונית חדשה תלוי במשך הזמן שעבר מאז קניית המכונית הקודמת. משך הזמן עד להגשת תביעת רכב חדשה בגין תאונה לחברת הביטוח תלוי במשך הזמן שעבר מאז התביעה הקודמת. משך הזמן עד שלקוח יעזוב את ספק הטלפון הסלולרי שלו ויעבור לחברה אחרת תלוי במשך הזמן שעבר מאז הלקוח רכש את השירות, ועוד דוגמאות רבות אחרות. הספרות המקצועית דנה בהרחבה במקרה הנומינלי (למשל, Kaplan and Meier 1958) ובמאמר זה אנו נדלג עליו.

2. משתנה תלוי נומרי (רציף)

כאמור, במקרה הנומרי (הרציף), המשתנה התלוי מבטא בדרך כלל זמן, והמטרה של ניתוח ההישרדות היא לאמוד את התפלגות הזמן עד לקרות האירוע (Time to event), שממנה ניתן לגזור מאפיינים שונים של תהליך ההישרדות. לדוגמה, ההסתברות שאדם ישרוד מעל t יחידות זמן לאחר קבלת טיפול מסוים, לחשב את תוחלת משך הזמן עד להחלמה ממחלה, להשוות את תוחלת משך הזמן ההחלמה ממחלה בין שתי קבוצות מטופלים, אחת שקיבלה טיפול תרופתי לעומת קבוצת ביקורת שקיבלה טיפול פלצבו, ועוד.

בסעיף זה נדון בקצרה במספר מאפיינים טיפוסיים של פונקציית ההישרדות. דיון רחב ניתן למצוא בספרות. סקירה ממצה של הנושא הכוללת שתי דוגמאות מתחום הרפואה, מופיעה במאמר Emmert-Streib et al. (2019).

נסמן:

T – משתנה הזמן

t – ערך ספציפי של משתנה הזמן

$F(t)$ – פונקציית הצפיפות של משתנה הזמן

$F(t) = \Pr(T \le t)$ – פונקציית ההתפלגות של משתנה הזמן

פונקציית ההישרדות $S(t)$ היא ההסתברות המשלימה ל-$F(t)$, כלומר ההסתברות שמשך הזמן שבו קורה האירוע גדול או שווה לזמן t. על סמך שיקולים הסתברותיים פשוטים:

$$S(t) = \Pr ob(Survival\ time \ge t) = \mathrm{PR}(T \ge t) = 1 - \Pr(T < t) = 1 - \mathrm{F}(t)$$

כלומר, פונקציית ההישרדות היא הפונקציה המשלימה (complement) של פונקציית ההתפלגות.

פונקציית ההישרדות היא לא אחידה אלא תלויה באופי של תהליך ההישרדות ובאופן שבו האירועים מתנהגים על פני זמן. מאפיינים אלה באים לידי ביטוי באמצעות פונקציית הסיכון (Hazard Function), $h(t)$, המתארת את ההסתברות שהאירוע המדובר יקרה בטווח זמן קצר dt (שנקרא לה ההסתברות ה"רגעית"), בהינתן שהאירוע לא קרה לפני זמן t. ובניסוח מתמטי:

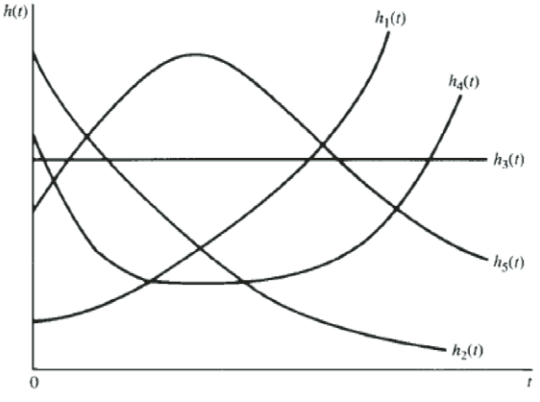
$$h(t) = \lim_{dt \to 0} \frac{\Pr\{t \le T < t + dt | T \ge t\}}{dt}$$

המונה בביטוי הזה אינו אלא ההסתברות המותנית (conditional probability) שהאירוע יקרה באינטרוול הזמן $(t\ ,\ t+dt)$ בהינתן שהאירוע לא קרה לפני כן, והמכנה הוא רוחב האינטרוול. על פי תורת ההסתברות, ההסתברות המותנית שווה ליחס בין ההסתברות המשותפת שהאירוע קורה באינטרוול $(t\ ,\ t+dt)$ בהינתן ש-$(T>t)$, ובין הסתברות התנאי. כאשר מדובר על אינטרוול שהוא קטן מאוד (כלומר dt שואף לאפס), נקבל על סמך שיקולים מתמטיים פשוטים:

$$h(t) = f(t)/S(t)$$

הביטוי h(t) מבטא את שיעור ההתרחשות (Rate of occurrence) של האירוע ליחידת זמן. נוסחה זו מאפשרת לעבור מפונקציית הסיכון לפונקציית ההישרדות, ולהיפך.

איור 2 מציג מספר פונקציות סיכון אפשריות:



איור 2: פונקציות סיכון

הפונקציה $h_1(t)$ מתארת פונקציית סיכון מונוטונית עולה
הפונקציה $h_2(t)$ מתארת פונקציית סיכון מונוטונית יורדת
הפונקציה $h_3(t)$ מתארת פונקציית סיכון קבועה
הפונקציה $h_4(t)$ מתארת פונקציית סיכון שיורדת בתחילה, עד שהיא מגיעה לערך מינימלי ואז חוזרת לעלות.
ולבסוף, הפונקציה$h_5(t)$ מתארת פונקציית סיכון שעולה בתחילה עד שהיא מגיעה לערך מקסימלי ואז יורדת בחזרה.

פונקציית הסיכון תלויה ביישום ומשתנה מיישום אחד למשנהו. לדוגמה, בענף ביטוחי הרכב, ההסתברות הרגעית שלקוח יגיש תביעת נזיקין לחברת הביטוח בגין תאונה היא לרוב קבועה, מכיוון שההסתברות להיות מעורב בתאונה היא בדרך כלל בלתי תלויה במשך הזמן שעבר מאז התאונה הקודמת. לעומת זאת, בענף של מכירות כלי רכב סביר להניח שפונקציית הסיכון לרכישת רכב חדש היא פונקציה עולה עם הזמן, שכן ככל שעבר יותר זמן מהמועד שבו אדם קנה את הרכב הנוכחי שלו, ההסתברות הרגעית שהוא ירכוש רכב חדש עולה.

הספרות המקצועית דנה במספר פונקציות סיכון רלוונטיות, שהפופולריות שבהן הן הפונקציה האקספוננציאלית (Exponential function), פונקציית וויבול (Weibull function), הפונקציה הלוג-לוגוסטית (Log-logistic function), הפונקציה הלוג-נורמלית (Log-normal function) ופונקציית גמה (Gamma function). לכל פונקציה יש את המאפיינים שלה וצורתה מתקבלת באמצעות הפרמטר של הפונקציה (שמסומנת בדרך כלל באות σ) שאותה אומדים על סמך נתוני הקלט. המאמר של Emmert-Streib et al., (2019) מציג דוגמאות של פונקציות סיכון אלה עבור מגוון רחב של פרמטרים.

עד לנקודה זו הנחנו שאוכלוסיית המדגם היא הומוגנית, כאשר משך החיים של כל תצפית מוגדר על ידי אותה פונקציית הישרדות $S(t)$. אולם ברור שזה לא המצב במציאות, שכן התצפיות באוכלוסייה שונות זו מזו ולכל אחת פונקציית הישרדות משלה התלויה במאפיינים של התצפית. בדומה למודלים ה"רגילים" של רגרסיה, נבטא את המאפיינים של כל תצפית באמצעות משתנים מסבירים (שבהקשר של פונקציית ההישרדות מכונים גם Covariates) על מנת לאמוד את פונקציית הסיכון הספציפית לכל תצפית. סדרה שלמה של מודלים, דמויי רגרסיה, המתמקדת באמידה של פונקציות סיכון עם משתנים מסבירים, פותחו על ידי המתמטיקאי הבריטי Cox. המודל הנפוץ ביותר הוא מודל הסיכון הפרופורציונלי (The Proportional Hazard Model), ובקיצור PHR. מודל הסיכון הפרופורציונלי הוא אחד המודלים היותר מורכבים בסטטיסטיקה ולא נפרט אותו כאן. דיון רחב בנושא מופיע בספרות המקצועית (Cox, 1972). סקירה מצוינת של נושא ניתוח ההישרדות גם עבור המקרה הבדיד וגם עבור המקרה הרציף מופיעה במאמר של Emmert-Streib et al., (2019). דיון נוסף על הנושא ניתן למצוא באתר ויקיפדיה.

3. יישומים של ניתוח הישרדות בתחום הניהול ומדעי החברה

מודלים בתחום הניהול ומדעי החברה הם בדרך כלל מודלים משני סוגים – מודלים של הסבר (Explanatory models) שמטרתם להסביר תופעות, ומודלים של חיזוי (prediction models) שמטרתם לחזות תופעות. דיון בהבדלים בין מודל הסבר למודל חיזוי מופיע אצל זהבי (2017). מודלים של רגרסיה, בין אם מדובר על מודלים של רגרסיה או מודלים של ניתוח הישרדות, מתבססים בדרך כלל על מדגם מייצג של האוכלוסייה, הידוע בשם מדגם הלמידה (Learning sample), שממנו מנסים להסיק על כלל האוכלוסייה. האופי וההרכב של מדגם הלמידה מכתיבים גם את סוג מודל הרגרסיה שבו יש להשתמש על מנת לנתח את הנתונים. השאלה שאנחנו צריכים לשאול היא האם אנו יכולים לחכות כדי לפעול ולקבל החלטות עד שכל הנתונים יהיו בידינו? במקרה של נתונים תלויי זמן התשובה היא לא! למשל, במחקר רפואי הבודק יעילות של תרופה מסוימת ומשתתפים בו כמה מאות אנשים, אנחנו לא יכולים לחכות עד שכל החולים שהשתתפו במחקר ימותו (או יחלימו) על מנת להעריך את יעילות התרופה. במחקר הבודק תקלות בייצור של מתקנים אלקטרוניים, אי אפשר לחכות עד שכל המתקנים יתקלקלו על מנת לאתר את מקור התקלה. במקרים כאלה אנחנו נאלצים לסגור את חלון התצפיות (Observations window) בנקודת זמן מסוימת מבלי שכל האירועים שהשתתפו במחקר הסתיימו. במצב כזה מדגם הלמידה מכיל תצפיות קטומות מימין, מפני שעבור תצפיות אלה אין לנו מידע על הזמן האמיתי עד שקורה האירוע. לא מן הנמנע שמדגם הלמידה הזה מכיל גם תצפיות קטומות משמאל, שהחלו עוד טרם החל תהליך איסוף התצפיות. כמובן שאין לשלול גם תצפיות עם קיטום מרווחי, הקטומות גם מימין וגם משמאל. ככלל, ניתן לומר שכאשר אנו עוסקים באירועים תלויי זמן, מדגם הלמידה מתבסס תמיד על חלון תצפיות שמתחיל בתחילת תקופת המחקר ומסתיים בסוף תקופת המחקר, אפילו לפני שכל האירועים במחקר התממשו, ולכן הוא כולל תצפיות קטומות מימין או משמאל, או תצפיות הקטומות גם מימין וגם משמאל. משך תקופת התצפיות משתנה בהתאם ליישום ולזמינות הנתונים. במונחים של ניתוח הישרדות, חלון התצפיות שקול לתקופת הניסוי.[1] גם אם משך חלון התצפיות הוא ארוך, כמו בעולם הרכב שהאורך שלו יכול לנוע בטווח של בין 5 ל-10 שנים, אנו עדים לתופעת הקיטום. למשל, אנשים שקנו את המכונית האחרונה שלהם לפני תחילת חלון התצפיות (קיטום משמאל), ואחרים שמכרו את הרכב שלהם לאחר תום חלון התצפיות (קיטום מימין). ובוודאי שבתקופה כה ארוכה ייתכנו גם תצפיות עם קיטום מרווחי, כלומר אנשים שקנו את הרכב הנוכחי שלהם לאחר תחילת חלון התצפיות והמירו אותו ברכב חדש עוד במהלך חלון התצפיות. כפי שתואר לעיל, במקרים של מדגמי למידה עם קיטום בנתונים, יש למודלים של ניתוח הישרדות יתרון על פני מודלים "מסורתיים" של רגרסיה בשל העובדה שיש להם את היכולת להתמודד עם תופעת הקיטום. בהמשך אנו נדגים תכונה זו של המודלים לניתוח הישרדות באמצעות דוגמה מספרית.

אומנם מרבית היישומים המסורתיים של ניתוח הישרדות הם בתחום הרפואה והבריאות, אבל אירועים תלויי זמן שכיחים מאוד גם בתחומים אחרים – ניהול, מדעי החברה, הנדסה ועוד – ולכן לניתוח הישרדות יש שימושים פוטנציאליים רבים גם בתחומים אלה. אלא שהדיון בספרות ביישומים של ניתוח הישרדות בתחומים שמעבר לתחום הרפואה הוא די דליל. נוסיף לכך את העובדה שהמודלים של ניתוח הישרדות הם בין המודלים היותר מורכבים בסטטיסטיקה ולכן הם פחות מוכרים בתעשייה. כתוצאה מכך, תהליך בניית המודלים, גם באירועים תלויי זמן, מתבסס לרוב על מודלים "רגילים" של רגרסיה במקום על מודלים של ניתוח הישרדות, מה שעשוי להביא למודלים מוטים ולהחלטות שגויות. גם הספרות המקצועית בתחום של ניתוח הישרדות לא עוזרת ליצור מודעות לפוטנציאל של המודלים לניתוח הישרדות בתחומי הניהול ומדעי החברה, מפני שהיא מתמקדת בעיקר ביישומים בתחום הרפואה והבריאות. למשל, Kleinback et al., (2011) משתמש בחמישה יישומים של ניתוח הישרדות כדי להסביר את הנושא, וכולם (חוץ מאחד) הם מתחום הרפואה. גם מקורות אחרים הדנים בניתוח הישרדות מתמקדים בעיקר ביישומים בתחום הבריאות.

כאמור, אירועים תלויי זמן לא ייחודיים רק לתחום הרפואה ונפוצים גם בתחומים אחרים, ובהם ניהול, שיווק, מדעי החברה, הנדסה ועוד, ולכן לניתוח הישרדות יש שימושים פוטנציאליים רבים גם מעבר לתחום הבריאות:

1 להלן נשתמש במושגים חלון (תקופת) התצפיות ותקופת הניסוי בתור מושגים נרדפים.

1. בנקאות – הפעלת אסטרטגיות שיווקיות בהתאם לערך הלקוח (Lifetime value)
2. ביטוח – משכי זמן של פיגורים בתשלום פרמיות ביטוח
3. משכנתאות – משך הזמן עד לגמר תשלומי המשכנתה
4. דיוור ישיר – משך הזמן עד לקנייה הבאה
5. קמעונאות – משך הזמן לאימוץ מוצרים חדשים
6. ייצור – אורך החיים של רכיבים
7. סקטור ציבורי – מרווחי זמן בין אירועים

בתחום השיווק וניתוח לקוחות, למודלים של ניתוח הישרדות יש ארבעה שימושים עיקריים:

1. תכנון עסקי – אפיון (Profiling) של לקוחות עם שיעורי הישרדות גבוהים יותר ונקיטת צעדים אסטרטגיים בהתאם (למשל, שימור לקוחות אלה באמצעות מתן תמריצים כספיים, מועדוני לקוחות, פעילויות חברתיות, ועוד).
2. חיזוי ערך הלקוח (Life time value) והתאמת הפעילות העסקית בהתאם לתחזיות של ערך הלקוח (למשל, טיפוח לקוחות עם ערך לקוח גבוה).
3. מעקב אחר רמת הפעילות העסקית של לקוחות והפעלת מנגנוני התערבות על מנת להגביר את הפעילות של הלקוחות הרווחיים (למשל, עידוד לקוחות שהקטינו את הפעילות הבנקאית שלהם להגדיל חזרה את נפח הפעילות העסקית שלהם).
4. בדיקת ההשפעה של מבצעי שיווק ושימור לקוחות על שיעורי ההישרדות של הלקוחות בארגון.

אף שלא לכולם יש סימוכין בספרות, להלן רשימה חלקית של שימושים אפשריים של ניתוח הישרדות בתחומים שונים:

- ניתוח לקוחות (Customer analytics) – נטישה של לקוחות ומעבר לחברה מתחרה
- ניתוח מוצרים (Product analytics) – משך הזמן לאימוץ (adoption) של מוצרים חדשים (למשל דור חדש של טלפונים סלולריים)
- ניהול משאבי אנוש – משך הזמן עד לקבלת קביעות (Tenure), משך הזמן לאיוש משרה פתוחה, משך זמן לקידום בעבודה
- שירות לקוחות – משך הזמן לטיפול בתקלות או בתלונות
- בנקאות – עמידה בתנאי תשלום הלוואות, יתרות של כרטיסי אשראי ומשכנתאות
- שיווק – משך הזמן של משק בית לקנית מוצר בר קיימא
- משחקים (Gaming) – משך זמן שמנויים מתמידים במשחק
- ניהול מלאי – משך הזמן לחידוש המלאי
- קרימינולוגיה – משך הזמן לשחרור על תנאי
- סוציולוגיה – משך הזמן בנישואים ראשונים
- תקשורת – משך זמן המנוי לעיתון או מגזין
- הנדסת אמינות – אורך החיים של מתקן אלקטרוני או מערכות אלקטרוניות
- ועוד

החדשות ה"טובות" הן שלאחרונה גדל העניין ביישום מודלים של ניתוח הישרדות גם בתחומי הניהול ומדעי החברה. להלן נביא שלוש דוגמאות מהספרות המקצועית שפורסמו לאחרונה:

המאמר של Gao et al., (2022) מנסה להסביר את התופעה של הורדת תרופות מהמדפים (Recall) עבור תרופה שעברה את כל האישורים הדרושים של FDA וכבר יצאה לשוק, וזאת כתגובה לתלונות צרכנים ברשתות החברתיות על בעיות בטיחותיות בשימוש בתרופה. המטרה של המחקר הזה הייתה לבדוק באופן אמפירי, בהקשר של תרופות, האם רשתות חברתיות יכולות להאיץ את התהליך של הורדת תרופות מהמדף (Recall). המחקר מתבסס על נתונים תלויי זמן בענף התרופות שסופקו על ידי ה-FDA, שכללו תרופות שהורדו המדף וכאלה שלא. המחברים ניסחו את בעיית המחקר באמצעות מודל רגרסיה. בשל העובדה שהתצפיות הן קטומות, שכן תופעת ה-Recall לכל תרופה קורית בזמנים שונים, מחברי המחקר העדיפו לאמוד את המקדמים של מודל הרגרסיה בכלים של ניתוח הישרדות, ובמקרה הזה במודל ניתוח הישרדות בדיד מכיוון שהנתונים נמדדו בנקודות זמן בדידות (חודשים). מדובר כאן על מודל הסבר שמנסה להסביר מהם המשתנים המשפיעים ביותר על תופעת ה-Recall, לתוצאות המחקר הזה יש משמעות רבה לתעשיית התרופות, שכן היא מחישה את התהליך של הורדת תרופות מזיקות מהמדף, ובכך מסייעת לצמצם את הנזק הבריאותי, הכספי והחברתי שנגרם על ידי תרופות עם תוצאות לוואי מזיקות ובלתי רצויות. לדוגמה, המחברים מציינים תרופה ללחץ דם שהעלתה את הסיכוי לחלות בסרטן, שהורדה מהמדפים מאוחר מדי, כמעט ארבע שנים לאחר שנכנסה לשוק, ולא לפני ש-60 מיליון פציינטים נחשפו אליה.

מחקר נוסף בתחום הניהול של (Garg et al., 2022), שעושה שימוש במודלים לניתוח הישרדות, עוסק בחיזוי סיכויי הכשל של פלטפורמות מקוונות לתשלומים (Online

payment platforms). הנתונים למחקר התקבלו משתי פלטפורמות סיניות מקוונות להלוואות על פני תקופה מסוימת (observations period). אומנם פלטפורמות מקוונות לתשלומים נחשבות ליציבות לאורך זמן, ועדיין יש מצב שפלטפורמות מקוונות ייכשלו ויפסיקו לפעול אחרי תום התקופה של איסוף הנתונים. כלומר, מדובר כאן על קיטום מימין של חלק מהנתונים. לכן בניתוח הנתונים מחברי המחקר השתמשו במודלים של ניתוח הישרדות ה"יודעים" לטפל טוב יותר בנתונים קטומים. גם כאן מדובר על מודל הסבר שמטרתו לבדוק מהם המשתנים המשפיעים על כשלים של פלטפורמות תשלומים מקוונות. המחברים סיווגו את המשתנים המשפיעים לארבע קטגוריות: מאפייני הפלטפורמה, ניהול סיכונים, מתחרים עסקיים, ומסרים מקוונים מפה לאוזן. בנוסף, המחברים איתרו מבין המשתנים המסבירים את המשתנים המשפיעים והחשובים באמצעות ערכי שפלי (Shapley values).[2] המחקר מספק למנהלי השיווק כלים ותובנות שמאפשרים להם להעריך טוב יותר את רמת הסיכון של פלטפורמות התשלומים, ונותן למשקיעים כלים לבחור את פלטפורמת התשלומים המועדפת עליהם.

ולבסוף, המאמר של Garg, S. et al., גם הוא משנת 2022, סוקר בהרחבה שימושים של למידת מכונה בתחום של ניהול משאבי אנוש, ומציין את הנושא של פיתוח קריירה (career development) שבו נעשה שימוש בניתוח הישרדות, עם דגש על שני יישומים: תחלופת עובדים (Turnover) וקידום קריירה (Career progression). המטרה היא להציע לעובדים הבכירים בארגון (Talents) מסלולי התקדמות שימנעו מהם לעבור לחברה מתחרה. ניתוח ההישרדות משמש כאן על מנת לחזות את משך הזמן שבו כל עובד "מבלה" בכל תפקיד או עובר מתפקיד לתפקיד. כמו ביישומים ברפואה, גם בבעיה זו החיזוי מוגבל לתקופת הניסוי בלבד. הנתונים לצורך חיזוי הם נתונים דמוגרפיים של העובד (גיל, מצב משפחתי וכדומה), היסטוריית התעסוקה של העובד (תאריך תחילת העבודה, תפקידים שמילא בעבר ומשך הזמן שהחזיק בכל תפקיד, תאריך פרישה וכדומה), ועוד. אלגוריתם מפורט לפתרון הבעיה, שמתבסס על נתונים שנאספו מחברת היי-טק גדולה במשך שנתיים, מוצג במאמר של Huayu Li et al., (2017).

2 ראו דיון על ערכי שפלי במאמר מקוון: https://christophm.github.io/interpretable-ml-book/shapley.html#fn43

4. הרחבה לחיזוי אנליטי

חיזוי אנליטי (Predictive analytics) הוא כיום אחד התחומים ה"חמים" ביותר של למידת מכונה. בעיות חיזוי נמצאות איתנו כבר שנות דור. מודל הרגרסיה הליניארית הוותיק הוא למעשה גם מודל חיזוי, אלא שהחיזוי מוגבל בדרך כלל רק עבור ערכים של המשתנים המסבירים הנמצאים בטווח הערכים של המשתנים במדגם הלמידה. המטרה של חיזוי אנליטי היא לחזות ערכים של אירועים עתידיים גם מעבר לטווח הערכים של נתוני הקלט. מדובר על תהליך שכולל חלוקת מדגם הלמידה למדגם אימון (Training sample) ומדגם תיקוף (Validation sample), בניית מודל על סמך מדגם האימון המכיל דוגמאות מהעבר, שממנו ניתן ללמוד על הקשר בין משתנה הפלט למשתני הקלט, תיקוף המודל על סמך מדגם התיקוף כדי לוודא שהוא מודל יציב ללא התאמת יתר (Overfitting), והפעלת המודל לצורך חיזוי המשתנה התלוי, בדרך כלל ההסתברות שהוא יקרה, גם עבור תצפיות חדשות שלא השתתפו במדגם הלמידה. תיאור מלא של תהליך החיזוי האנליטי מופיע אצל זהבי (2017).

אולם מודל ההישרדות שבו דנו לעיל לא מתאים כדי לחזות הסתברויות של אירועים עתידיים. סיבה אחת היא שמודלים של הישרדות מוגבלים רק לאמידת הסיכויים של זמני הישרדות בתקופת הניסוי, בדרך כלל, אם כי לא תמיד, ברזולוציה של שנים, לגבי התצפיות שהשתתפו במדגם הלמידה. הסיבה השנייה היא שמודלים של חיזוי אנליטי לא מתבססים על התפלגויות של זמני הישרדות אלא על אירועים בפועל (למשל, קניות או אי קניות של רכבים). לעומת זאת, בבעיות חיזוי אנליטי, עבור אירועים תלויי זמן, המטרה היא לחזות את הסיכויים שהאירוע יקרה בזמן עתידי מעבר לתקופת הניסוי, וזאת כתלות במועד שעבר מאז שקרה האירוע הקודם. כלומר, מדובר כאן בהסתברויות מותנות. לדוגמה, בענף הרכב אנו מבקשים לחזות עבור כל לקוח בבסיס הנתונים, מהי ההסתברות המותנית שהוא יקנה רכב חדש גם מעבר לתקופת הניסוי, כתלות במשך הזמן שעבר מאז קניית הרכב הקודם. בבעיות נטישה (Churning) נרצה לאמוד את ההסתברות המותנית שהלקוח יעזוב את הספק שלו (למשל, חברה סלולרית מסוימת) ויעבור לספק אחר, כתלות ב"זמן" שלו בספק המקורי. והדוגמאות הן רבות.

על מנת להתמודד עם בעיות חיזוי עתידיות גם בתהליכים ניהוליים תלויי זמן, נגדיר להלן מודל שמשלב בין מודלים של

הישרדות ומודלים של חיזוי אנליטי. הרעיון הוא להשתמש במודלים מהתחום של ניתוח הישרדות על מנת לאמוד את הפרמטרים של המודל הסטטיסטי, וכך גם להתמודד עם תופעות הקיטום שהמודלים ה"מסורתיים" ברגרסיה לא מסוגלים לטפל בהן. לאחר מכן להשתמש בעקרונות של גישות לחיזוי אנליטי מבוססי רגרסיה, שמשתמשים באומדים שהתקבלו ממודל ההישרדות על מנת לבנות מודל חיזוי, לתקף אותו, ולהפעיל את המודל לצורך חיזוי האירועים על תצפיות חדשות. פירוט של התהליך ודוגמה מעשית יופיעו בהמשך.

הגישה הנ"ל יושמה בחבילת התוכנה GainSmarts (Levin and Zahavi, 2005). לצורך כך השתמשנו במודל דמוי רגרסיה, אחת מהפונקציות במשפחת הפונקציות של מודל הסיכון הפרופורציונלי של Cox. המודל שבו השתמשנו לצורך הרחבה של גישת החיזוי האנליטי לתהליכים על פני זמן הוא מודל דמוי רגרסיה מהסוג:

$$Log(T) = B' * X + \sigma * W$$

כאשר:

T – משתנה מקרי המציין את הזמן.

W – פונקציית הסיכון, אחת מפונקציות הסיכון שתוארו לעיל, שאותה בוחרים מראש על פי האופי של תהליך הסיכון.

σ – פרמטר של פונקציית הסיכון שקובע את ההתנהגות של פונקציית הסיכון, האם היא קבועה, עולה או יורדת כפונקציה של הזמן, או עולה עם הזמן עד שהיא מגיעה לשיא ואז יורדת עם הזמן. σ היא פרמטר שיש לאמוד על סמך התצפיות במדגם הלמידה.

בדומה למודל רגרסיה ליניארית "רגיל", האיבר $B' * X$ [3] מבטא את התוחלת של המודל, והאיבר $\sigma * W$ את המרכיב האקראי שלו, ולכן זהו מודל דמוי רגרסיה.

במודל המשולב הישרדות/רגרסיה חילקנו את ציר הזמן למספר משכי זמן, כמתואר באיור 3:

תקופת האימון (Training period) – תקופת הזמן לצורך אימון המודל. מועד ההתחלה שלה הוא תחילת תקופת הניסוי ונקודת הסיום שלה היא נקודת הזמן M0. תקופת האימון היא התקופה שעל פיה מגדירים את כל המשתנים המסבירים במודל כפי שהם נמדדים בנקודת הזמן M0. במילים אחרות, הקובץ המשמש לאימון המערכת משקף את תמונת המצב (Snapshot) של המשתנים המסבירים בנקודת הסיום של תקופת האימון.

תקופת היעד (Target period) – תקופת הזמן לצורך הגדרת הבחירה (choice) של הלקוח (המשתנה התלוי). אם האירוע (למשל קניית רכב) קרה בפרק בזמן הזה, משתנה ה-Choice מקבל את הערך 1, אחרת הוא מקבל את הערך 0. תקופת היעד מתחילה במועד הסיום של תקופת האימון, כלומר נקודת הזמן M0, ומסתיימת בנקודת הזמן M1 שמגדירה את התאריך של "היום" (שהוא גם התאריך הנוכחי שבו מריצים את המודל). נקודת הזמן M1 היא גם מועד הסיום של תקופת הניסוי. בדרך כלל תקופת היעד היא תקופה קצרה יחסית, למשל שנה או שנתיים בעולם הרכב. ולכן בשלב זה אנו מניחים שבתקופת היעד קורה בדרך כלל רק אירוע אחד, שכן לא סביר שבתקופה כה קצרה אדם יקנה יותר מרכב אחד או כל מוצר בר-קיימא אחר.

תקופת החיזוי (Prediction period) – התקופה מעבר לתקופת הניסוי שלוקחת חלק בתהליך החיזוי עבור תצפיות חדשות. זהו פרק הזמן מסוף תקופת הניסוי (M1) ובין סוף תקופת החיזוי (M2). משך תקופת החיזוי נקבע על ידי המשתמש.

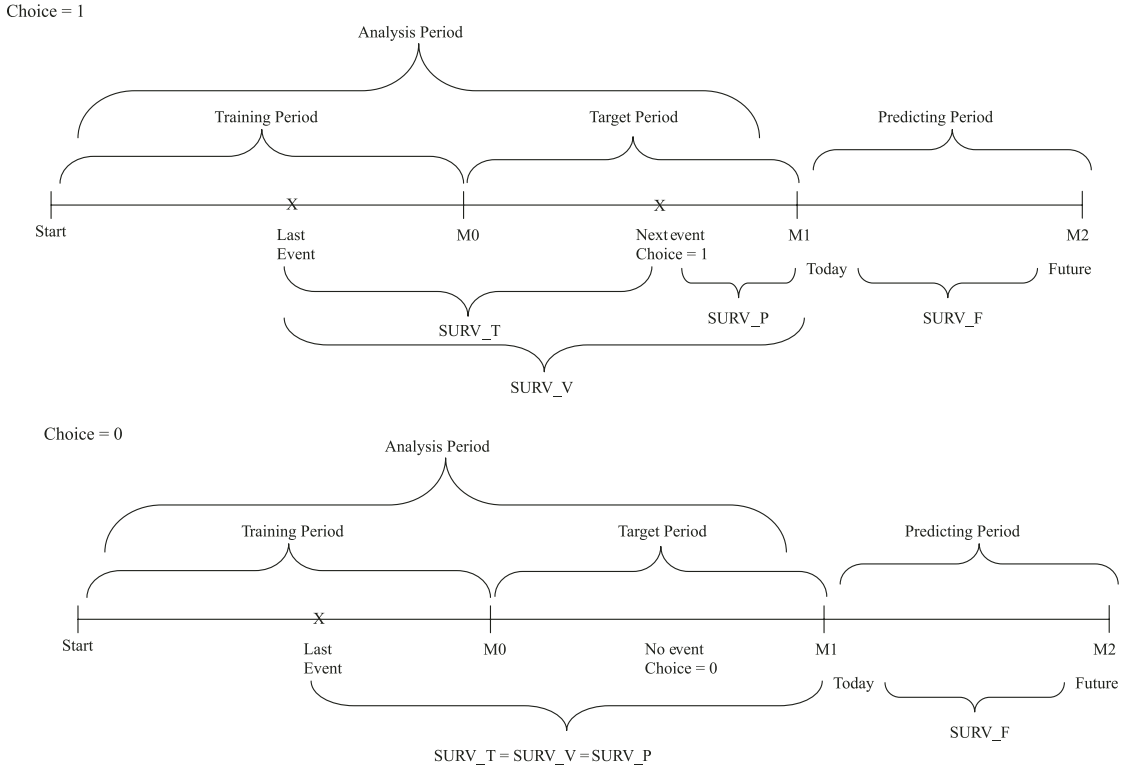
לצורך בניית המודל המשולב הישרדות/רגרסיה אנו מגדירים ארבעה משתני זמן, כמפורט באיור 3.

SURV_T: הזמן שבין האירוע האחרון בתקופת האימון והאירוע הבא בתקופת היעד. או אם לא קרה אירוע בתקופת היעד, הזמן עד סוף תקופת היעד. התקופה הזו משתתפת בתהליך בניית המודל.

SURV_V: הזמן שבין האירוע האחרון בתקופת האימון וסוף תקופת היעד, ללא תלות אם קרה או לא קרה אירוע בתקופת היעד. תקופה זו לוקחת חלק בתהליך בדיקת האיכות של מודל החיזוי ותיקוף שלו.

SURV_P: התקופה הקודמת לתקופת החיזוי – הזמן שבין האירוע האחרון בתקופת היעד וסוף תקופת היעד. הזמן הזה משמש בתהליך החיזוי עבור הלקוחות החדשים.

3 נזכיר רק שהביטוי $B' * X$ מבטא את המכפלה הסקלארית של וקטור המקדמים B (וקטור שורה) בווקטור המשתנים X (וקטור עמודה), כלומר $b_1 * X_1 + b_2 * X_2 + , \dots , b_j * X_j$



איור 3: חלוקה לתקופות זמן

SURV_F: תקופת החיזוי שמתחילה נקודת הזמן M1 ("היום") ועד לנקודת הסיום M2. תקופת החיזוי היא התקופה שעבורה אנחנו רוצים לחזות את מועד האירוע הבא עבור הלקוחות החדשים. אורכה של תקופת החיזוי נקבע על ידי המשתמש.

תהליך החיזוי האנליטי הוא תהליך רב שלבי שבו משתתפים מספר קבצים (קובץ האימון, קובץ התיקוף וקובץ הלקוחות החדשים), ומספר תקופות זמן (תקופת הניסוי, תקופת האימון, תקופת התיקוף ותקופת החיזוי), כמפורט להלן:

- **הגדרה של פונקציית הסיכון**: תהליך החיזוי מתחיל עם ההגדרה של פונקציית הסיכון המאפיינת את הבעיה ה"תורנית". למשל, אם מדובר במבצע שיווק לרכב חדש, סביר שפונקציית הסיכון היא מונוטונית עולה, שמשמעותה שככל שהלקוח מחזיק ברכב הנוכחי שלו משך זמן ארוך יותר, כך ההסתברות שהוא ירכוש רכב חדש בתקופה הבאה גדולה יותר. בבעיות של תביעות על נזקים ברכבים כתוצאה מתאונה, סביר להניח שפונקציית הסיכון היא קבועה ומשמעותה שההסתברות להגיש תביעת נזיקין לרכב היא בדרך כלל בלתי תלויה במועד שעבר מאז הגשת התביעה הקודמת. זוהי האחריות של המשתמש להגדיר את פונקציית הסיכון עבור בעיית ההחלטה ה"תורנית", וזאת על סמך ההיכרות שלו עם הבעיה והמאפיינים שלה.

- **אמידת פונקציית ההישרדות:** בהינתן פונקציית הסיכון, השלב הבא הוא אמידת פונקציית ההישרדות. האמידה מבוססת על קובץ האימון על פני תקופת הניסוי. המשתנה התלוי הוא משתנה הזמן (**SURV_T** באיור 3). המשתנים המסבירים הם המאפיינים של התצפיות כפי שהם נמדדים בזמן M0, כגון גיל, מצב משפחתי, המועד שבו קרה האירוע הנוכחי, ועוד שורה ארוכה של משתנים הרלוונטיים לבעיה ה"תורנית". במודל המשולב אנו אומדים את הפרמטרים (וקטור המקדמים של המשתנים המסבירים ואת סטית התקן σ שהיא הפרמטר של פונקציית הסיכון) בכלים הסטטיסטיים של ניתוח ההישרדות שמתחשבים גם בתופעת הקיטום של האירועים על פני זמן. האמידה של הפרמטרים של המודל בתוכנת GainSmarts נעשתה באמצעות הפרוצדורה LIFEREG של SAS.

- **בדיקת איכות החיזוי של מודל ההישרדות:** בשלב זה נכנסים לתמונה העקרונות המבוססים על הרגרסיה הלוגיסטית. בדיקת איכות החיזוי נעשית על בסיס קובץ האימון. לצורך זה יש להפעיל את מודל ההישרדות שנאמד בשלב הקודם על קובץ האימון, כדי לחשב את ההסתברויות שהאירוע שבו מדובר קרה בתקופה SURV_V באיור 3 (תהליך הנקרא Scoring, ציינון בעברית). נציין שמשך תקופת האימון משתנה מתצפית לתצפית בהתאם למועד שבו קרה האירוע האחרון בתקופת היעד. ההסתברויות המתקבלות מאפשרות לסכם את תוצאות המודל בטבלאות רווחים (Gain charts), ברמה של עשירונים (או כל אחוזון אחר), שמהם ניתן לגזור מדדי ביצוע שונים על איכות המודל ויכולת החיזוי שלו. מודל חיזוי "טוב" הוא מודל שעושה הבחנה טובה בין ה"קונים" ל"לא קונים".

- **תיקוף מודל החיזוי:** בהינתן שמדדי הביצוע על קובץ האימון מצביעים על מודל חיזוי "טוב", השלב הבא הוא תיקוף המודל שנועד לוודא שהמודל הוא גם בעל יכולת הכללה (Generalization) וניתן ליישם אותו גם על תצפיות חדשות מעבר למדגם הלמידה. על מנת לתקף את המודל יש לחזור על תהליך הציינון, אבל הפעם על קובץ התיקוף, ולגזור את אותם מדדי ביצוע לאלה שחישבנו עבור קובץ האימון. גם כאן תקופת התיקוף היא התקופה SURV_V. מאחר שקובץ התיקוף מכיל גם את האירועים בפועל, תיקוף המודל נעשה באמצעות השוואה של הביצועים בפועל של המודל בין קובץ האימון לקובץ התיקוף.

- **חיזוי עבור לקוחות חדשים:** לאחר שמתקפים את המודל ומוודאים שהוא ניתן להכללה, ניתן להפעיל אותו על התצפיות החדשות על מנת לחזות את ההסתברות המותנית שיקרה אירוע בתקופת החיזוי SURV_F, כלומר בתקופה בין M1 ל-M2, בהינתן משך הזמן שעבר מאז האירוע האחרון בתקופה שקדמה לתקופת החיזוי, SURV_P. על פי תורת ההסתברות, ההסתברות המותנית שיקרה אירוע בתקופת זמן עתידית Q, נתון שלא קרה שום אירוע קודם בתקופת הזמן L, היא:

$$Prob(t) < L + Q \mid t > L) = \frac{Prob(L < t < L + Q)}{Prob(t > L)} = 1 - \frac{S(L + Q)}{S(L)}$$

ובמקרה שלנו אנו מחפשים את ההסתברות המותנית:

$$Prob(t < \text{SURV\_P} + \text{SURV\_F} \mid t >= \text{SURV\_P})$$

נציין שקובץ הנתונים שעליו עושים את התחזית הוא הקובץ של הלקוחות החדשים שמתייחס לנקודת הזמן M1, כלומר משקף את תמונת המצב של התצפיות "היום".

בשלב החיזוי ניתן לחשב גם פרמטרים נוספים. פרמטר חשוב הוא החציון (Median) של זמן ההישרדות עבור כל תצפית. החציון הוא הערך של הזמן t שעבורו ההסתברות המותנית שהאירוע קרה לפני זמן t הוא 50%, וההסתברות המותנית שהאירוע יקרה לאחר זמן t הוא 50%. החציון מאפשר לבטא את פונקציית ההישרדות של התצפית באמצעות פרמטר מייצג אחד, מה שמאפשר מצד אחד לשלב אותו בחישובים של מדדים ביצועיים, כגון ערך לקוח, ומצד שני להשוות זמני הישרדות בין תצפיות שונות.

5. דוגמה מעשית

לבסוף, נדגים את המודל המשולב הישרדות/רגרסיה על בעיה מתחום הרכב. חברות הרכב הגדולות נוהגות בדרך כלל ללוות מבצעים להשקת רכב חדש, או מודל מתקדם יותר של רכב קיים, במבצעים של דיוור ישיר על מנת לעורר יותר מודעות (Awareness) לרכב החדש בקרב הלקוחות שלהם, להדגיש את התכונות של הרכב החדש מול המתחרים, ולעודד את הלקוחות ליצור קשר עם סוכני המכירות של החברה ולהגיע לאולמות התצוגה. בדרך כלל הפנייה נעשית רק ללקוחות השייכים לפלח (סגמנט) המתאים לרכב החדש. לדוגמה, בהשקה של רכב מסחרי (למשל טנדר מסוג F150 של חברת Ford), סגמנט הרכב המתאים הוא משקי בית בענף החקלאות, המסחר והתעשייה הקלה, ו/או משפחות שגרות באזורים כפריים ובעיירות קטנות. למרות שאי אפשר לשלול את האפשרות שגם תושבי ניו יורק או תל אביב יקנו רכב מסחרי, הסבירות לכך היא הרבה יותר קטנה, ולכן תושבי ערים גדולות לא שייכים בדרך כלל לסגמנט של קוני רכב מסחרי. מכל מקום, גם לאחר סינון הלקוחות השייכים לסגמנט הרכב המתאים, עדיין אין טעם לפנות לכל הלקוחות בסגמנט, שרובם מין הסתם לא מתכוונים להגיב להצעה, אלא למקד את הפנייה רק ללקוחות עם הסבירות הגבוהה ביותר לרכוש את הרכב החדש בתקופה הקרובה. כאן נכנס לתמונה הנושא של חיזוי אנליטי. מאחר שקניית רכב היא מאורע תלוי זמן, ובשל העובדה שבשוק הרכב הסבירות של לקוח לקנות רכב חדש תלוי במועד שבו קנה את הרכב הנוכחי

שלו, יש עדיפות להשתמש בניתוח הישרדות על מנת לאתר את הלקוחות האלה.

מסיבות של סודיות מסחרית אנו מנועים להשתמש בקובצי נתונים אמיתיים על מנת להדגים את הנושא. לכן ייצרנו קובץ "סינתטי", GSM ("GainSmarts Motors"), שמדמה את הנתונים בענף הרכב באמצעות סימולציה, ומכיל משתנים שהם אופייניים לענף הרכב, ברמה של משק הבית (household), ובהם: הגיל של בעל הבית, סוג הבית – כפרי/עירוני, הכנסה שנתית משוערת של משק הבית, משך הזמן בשנים שהמשפחה גרה בבית, ומספר הנפשות המתגוררות בבית. בנוסף, משתנים הקשורים לרכבים של משק הבית, כגון תאריך הקנייה האחרון של רכב (אם היה) בתקופת האימון, תאריך הקנייה האחרון של רכב (אם היה) בתקופת היעד, מספר הרכבים הכולל שנקנו על ידי משק הבית בחמש השנים האחרונות ובעשר השנים האחרונות, היכן משק הבית מבטח את רכבו, האם משק הבית נענה לסקרי שביעות רצון של חברת הרכב, ועוד. כמו כן, קובץ הלמידה כלל גם את כל ארבעת משתני הזמן שפורטו לעיל: SURV_T, SURV_V, SURV_P, ו- SURV_F. את אורך תקופת החיזוי, SURV_F, עבור לקוחות חדשים קבענו על 12 חודשים בלי קשר להיסטוריה של רכישות הרכב של משק הבית. הקובץ מנה 50,000 משקי בית המהווים מדגם מקרי של כלל אוכלוסיית הסגמט ומשקפים את ההתפלגות באוכלוסייה.

לצורך בניית המודל, חילקנו את הקובץ באופן מקרי לשני קבצים בלתי תלויים – קובץ אימון שכלל 2/3 של הקובץ והכיל 33,533 משקי בית, מהם 10,013 תגובות (שיעור תגובה של 29.86%), וקובץ תיקוף שכלל את ה-1/3 הנותר של הקובץ ומנה 16,647 משקי בית, מהם 4,911 תגובות (שיעור תגובה של 29.50%, בדומה לשיעור התגובה של קובץ האימון). בשני הקבצים התגובות מתייחסות לקניות רכב בתקופת היעד. קובץ האימון שימש לצורך בניית המודל, וקובץ התיקוף לצורך תיקוף המודל.

כאמור לעיל, אמידת הפרמטרים של המודל נעשתה על סמך מודל ההישרדות. המודל הסופי כלל 14 משתנים מסבירים (מתוך 25 המשתנים המקוריים בקובץ הקלט). פונקציית הסיכון שנבחרה היא פונקציית Weibull. הערך של האומד של σ מתוך המודל הוא 0.0671, ערך חיובי שעבור פונקציית Weibull מבטא פונקציית סיכון עולה (כלומר שההסתברות לקנות רכב חדש גדלה ככל שמשך הזמן מאז קניית הרכב הקודם גדול יותר).

משלל הדוחות של המערכת, נציג כאן שתי דוחות מרכזיים: טבלת הרווחים (Gains chart) עבור קובץ התיקוף, וטבלה שמשווה מספר מדדים מקובלים בעולם השיווק, ברמה של עשירונים, בין קובץ האימון וקובץ התיקוף.

הדרך המקובלת ביותר להציג את תוצאות מודל החיזוי, שממנו גם ניתן ללמוד על טיב החיזוי של המודל, הוא באמצעות טבלת רווחים (Gains table). טבלת רווחים משווה למעשה את תוצאות החיזוי מול התוצאות בפועל, בדרך כלל ברמה של עשירונים, בסדר יורד של ה"ציונים" של התצפיות בקובץ. נזכיר שבמודל המשולב הישרדות/רגרסיה, ה"ציונים" של הלקוחות מבטאים את ההסתברויות שהאירוע (קניית רכב חדש) יקרה בתקופת התיקוף (SURV_V באיור 3). על מנת לקבל את טבלת הרווחים ממיינים את הלקוחות בסדר יורד של התחזיות של ההסתברויות, מהציון הגבוה לנמוך, ומסכמים את הנתונים (קניות רכב) ברמה של עשירונים או בכל אחוזון אחר. במודל "טוב", סידור הלקוחות בסדר יורד של ההסתברויות מציב את הלקוחות ה"טובים" עם הסבירות הגבוהה לקנות רכב חדש בראש הרשימה ואת הלקוחות הפחות "טובים" בתחתית הרשימה, מה שמאפשר להעריך את טיב החיזוי.

ניתן ליצור את טבלת הרווחים על קבצים שונים. במאמר זה נציג את טבלת הרווחים עבור קובץ התיקוף (טבלה 1), שכן קובץ זה לא משתתף בתהליך בניית המודל ולכן הוא מייצג את "הלקוחות החדשים" שמהם אנחנו רוצים לבחור את הלקוחות שישתתפו במבצע השיווק המלא. עם זאת, מאחר שקובץ התיקוף מכיל גם את התוצאות בפועל, אפשר "על הדרך" להשוות את תוצאות החיזוי לתוצאות בפועל ומכאן ללמוד על איכות המודל והאם ניתן להשתמש בו לצורך בחירת הלקוחות החדשים שייקחו חלק במבצע השיווק.

מתוך הטבלה קל לראות את טיב התחזית של המודל באמצעות מספר מדדים. לדוגמה, מספר התגובות (קניות רכב) בפועל (Responses) הוא פונקציה מונוטונית יורדת כשנעים על פני העשירונים, מהעשירון העליון לתחתון, מה שאומר שהמודל אכן מצליח למקם את הלקוחות ה"טובים" בעשירונים העליונים של טבלת הרווחים. בעשירון העליון, וגם בזה שאחריו, מספר המגיבים בפועל הוא 1,647, שהם 100% מהלקוחות בעשירונים האלה. נשים לב שגם התחזיות של ההסתברויות

טבלה 1: טבלת הרווחים עבור המודל המשולב הישרדות/רגרסיה על קובץ התיקוף.

Resp. Prob.	Customers	% Customers	Responses	% Response	Actual Response Rate (%)	%Resp./ %Cust.	Pred. Resp.	Pred. RR(%)
99.83	1647	10.0	1647	35.5	100.00	3.4	1646	99.97
93.25	1647	10.0	1647	35.5	100.00	3.4	1616	98.12
17.78	1647	10.0	809	16.5	49.12	1.6	629	38.10
12.38	1646	10.0	293	6.0	17.80	0.6	242	14.71
9.62	1647	10.0	191	3.9	11.60	0.4	180	10.93
7.57	1647	10.0	134	2.7	8.14	0.3	140	8.49
6.02	1646	10.0	97	2.0	5.89	0.2	112	6.77
4.69	1647	10.0	51	1.0	3.10	0.1	88	5.33
3.34	1647	10.0	30	0.6	1.82	0.1	66	4.02
0.35	1646	10.0	12	0.0	0.73	0.0	38	2.30

בעשירונים אלה קרובות ל-100%! ככל הנראה מדובר כאן בלקוחות שקנו את הרכב הנוכחי שלהם לפני זמן רב ולכן הם "בשלים" להחליף את רכבם הישן ברכב חדש. ואומנם, ככל שיורדים ברמה של העשירונים, מספר התגובות (קניות רכב חדש) הולך ויורד. למשל, בעשירון התחתון המודל שלנו חוזה רק 12 קניות רכב, המהוות רק 0.73% מהלקוחות של העשירון. כאן ההסבר הוא שמדובר על לקוחות שקנו את רכבם הנוכחי לאחרונה, ולכן בשלב זה אין להם עניין רב לקנות רכב חדש.

תופעה דומה קורית גם כשמסתכלים על תחזית מספר התגובות של המודל. כך למשל, המודל צופה 1,646 מגיבים (Pred. Resp.) בעשירון העליון, כלומר שיעור תגובה חזוי (Pred. RR(%)) של כמעט 100%, לעומת 38 מגיבים בלבד בעשירון התחתון עם שיעור תגובה חזוי של כ-2.30%. מכלל הלקוחות בעשירון הזה. מדד נוסף לאיכות החיזוי של המודל הוא ההשוואה בין מספר המגיבים בפועל (Responses) לתחזית מספר המגיבים (Pred. Resp.). גם כאן ישנה התאמה טובה, אם כי לא מלאה, בין תחזיות המודל למספר המגיבים בפועל. כך למשל, מספר המגיבים בפועל בעשירון העליון הוא 1,647, מספר כמעט זהה למספר המגיבים החזוי. כך גם בעשירון השני. בעשירונים הבאים אנו עדים להבדלים מעט יותר גדולים בין מספר המגיבים בפועל לבין מספר המגיבים החזוי. אי אפשר כמובן לצפות לזהות מלאה ברמת העשירונים בין מספר המגיבים בפועל לתחזית מספר המגיבים. אבל בתור כלל "אצבע", סטייה של התחזית בסדר גודל של 5%–10% נראית סטייה סבירה.

טבלה 2 מציגה את תוצאות התיקוף של המודל הנ"ל באמצעות השוואה של תוצאות המודל בין קובץ האימון לקובץ התיקוף. הטבלה מציגה מספר מדדי ביצוע של המודל, זה בצד זה, כפונקציה של העשירונים, כאשר TRN מסמן את קובץ האימון (Training) ו-VLD את קובץ האימות (Validation). בכל המדדים ישנה התאמה טובה בין התוצאות בפועל בקובץ האימון לקובץ התיקוף, מה שמעיד על כך שמדובר במודל יציב ובר הכללה שניתן ליישם אותו לחיזוי התגובה של משקי בית חדשים.

נציין שפונקציית Weibull שבה השתמשנו מתאימה למקרה שבו מדובר על קניות רכב מאותו מותג (למשל Ford), שכן סביר להניח שככל שמשך הזמן מהקנייה האחרונה ארוך יותר, כך גם הסבירות לקנות רכב חדש מאותו המותג גדולה יותר. אבל כידוע לקוחות שקונים רכב חדש לא קונים בהכרח רכב מאותו מותג אלא גם ממותגים אחרים (שלגביהם אין לנו אינפורמציה). במקרה כזה סביר להניח שבתחילה הסבירות שלקוח שיש לו כבר רכב מאותו מותג (Ford בדוגמה הנ"ל) תגדל עם הזמן, אבל אם לאחר תקופת זמן מסוימת הלקוח עדיין לא קנה רכב של Ford, ההסתברות הרגעית שהוא יקנה

טבלה 2: מדדי ביצוע בין קובץ האימון לבין קובץ התיקוף עבור המודל המשולב הישרדות/רגרסיה

Decile	% Resp. TRN	% Resp. VLD	%Resp./%Cust. TRN	%Resp./%Cust. VLD	Actual RR(%) TRN	Actual RR(%) VLD	Pred. RR(%) TRN	Pred. RR(%) VLD
1	33.5	33.5	3.3	3.4	100.00	100.00	99.96	99.97
2	33.5	33.5	3.3	3.4	100.00	100.00	98.07	98.12
3	16.9	16.5	1.7	1.6	50.43	49.12	39.45	38.10
4	6.1	6.0	0.6	0.6	18.19	17.80	14.89	14.71
5	3.8	3.9	0.4	0.4	11.48	11.60	10.97	10.93
6	2.7	2.7	0.3	0.3	8.08	8.14	8.61	8.49
7	1.7	2.0	0.2	0.2	5.04	5.89	6.85	6.77
8	0.9	1.0	0.1	0.1	2.83	3.10	5.38	5.33
9	0.5	0.6	0.0	0.1	1.40	1.82	4.03	4.02
10	0.4	0.2	0.0	0.0	1.13	0.73	2.30	2.30

רכב של Ford הולכת ויורדת עם הזמן מפני שסביר להניח שהוא קנה רכב של מותג אחר. במקרה הזה חלופה סבירה יותר עבור פונקציית הסיכון היא הפונקציה הלוג-לוגיסטית, מכיוון שזוהי פונקציה שבתחילה עולה עם הזמן עד שהיא מגיעה למקסימום, ואז יורדת כפונקציה של הזמן (דוגמה טובה של פונקציה לוג-לוגיסטית היא הפונקציה $h_5(t)$ באיור 2). דיון זה רק ממחיש את החשיבות של הבחירה הנכונה של פונקציית הסיכון שמתארת את התנהגות האירוע. לא מן הנמנע שלעיתים יש צורך להריץ את הבעיה ה"תורנית" על מספר פונקציות סיכון, ואז לבחור את הפונקציה התאימה ביותר על סמך התוצאות של המודלים השונים.

לבסוף, כדי להדגים את הבעייתיות בשימוש שגוי במודלי חיזוי על מסדי נתונים עם נתונים קטומים, הרצנו את בעיית הרכב הנ"ל, גם עם מודל רגרסיה לוגיסטית "רגיל", על אותו קובץ נתונים. המשתנה התלוי היה קניות הרכב בתקופת היעד, וכל המשתנים האחרים בתקופת האימון, כולל משתני הזמן, שימשו כמשתנים מסבירים. על מנת להשוות "תפוחים לתפוחים" הרצנו את מודל הרגרסיה עם אותם פרמטרים של המודל המשולב, אותו תהליך של העיבוד המקדים (Preprocessing) ואותו תהליך של בחירת המשתנים המסבירים למודל (Feature selection). טבלה 3 מציגה את טבלת הרווחים על קובץ התיקוף.

טבלה 3: טבלת הרווחים עבור מודל הרגרסיה הלוגיסטית על קובץ התיקוף.

Resp. Prob.	Customers.	% Customers	Responses	% Response	Actual Response Rate (%)	%Resp./ %Cust.	Pred. Resp.	Pred. RR(%)
77.04	1647	10.0	1469	29.9	89.19	3.0	1484	90.12
55.04	1647	10.0	1077	21.9	65.39	2.2	1095	66.47
38.98	1647	10.0	728	14.8	44.20	1.5	761	46.18
27.47	1646	10.0	604	12.3	36.70	1.2	545	33.08
19.10	1647	10.0	390	7.9	23.68	0.8	375	22.76
12.75	1647	10.0	276	5.6	16.76	0.6	258	15.65
9.21	1646	10.0	186	3.8	11.30	0.4	184	11.16
5.45	1647	10.0	109	2.2	6.62	0.2	120	7.30
3.16	1647	10.0	58	1.2	3.52	0.1	70	4.26
0.66	1646	10.0	14	0.3	0.85	0.0	31	1.90

טבלה 4: השוואת מספר התגובות המצטברות על קובץ התיקוף, ברמה של עשירונים, בין המודל המשולב הישרדות/רגרסיה לבין מודל רגרסיה לוגיסטית "רגיל"

עשירון	1	2	3	4	5	6	7	8	9	10
מודל משולב	1647	3294	4103	4396	4587	4721	4818	4869	4899	4911
רגרסיה	1469	2546	3274	3878	4268	4544	4730	4839	4897	4911
הפרש	178	748	829	518	319	177	88	30	2	0

על פניו, מדובר על מודל רגרסיה "טוב" שמצליח למקם את הלקוחות ה"טובים" בעשירונים העליונים של טבלת הרווחים. בנוסף גם וידאנו שמדובר על מודל יציב ללא התאמת יתר. השוואה בין טבלה 1 לטבלה 3 מצביעה על יתרון בולט של המודל המשולב הישרדות/רגרסיה על פני מודל הרגרסיה הלוגיסטית מבחינת ההפרדה בין הלקוחות הטובים לאלה הפחות טובים. טבלה 4 מפרטת את מספר הקונים (Responses) המצטבר בין שני המודלים, שממנו מסתבר שהמודל המשולב הישרדות/רגרסיה מצליח "לתפוס" יותר קוני רכב מאשר מודל הרגרסיה הלוגיסטית בכל רמת עשירון, ובמיוחד בעשירונים העליונים. כך למשל, המודל המשולב "תופס" 829 יותר קונים בעשירון השלישי, המהווים כ-17% מכלל התגובות(!), 518 תגובות קונים בעשירון הרביעי (10.5% מכלל התגובות), וכך הלאה. מאחר שתהליך החיזוי האנליטי היה זהה בשני המקרים, ההבדלים בתוצאות החיזוי אלה נובעות כנראה מההטיה הנוצרת בגין השימוש במודל רגרסיה לוגיסטית על מסד נתונים קטום. תוצאות אלה רק ממחישות את היתרונות של שימוש במודלים של ניתוח הישרדות כשמדובר על אירועים תלויי זמן.

סיכום

ניתוח הישרדות (survival analysis) עוסק באמידה של פונקציית ההתפלגות של משך הזמן עד שקורה אירוע (זמן ההישרדות) עבור נתונים קטומים. בהינתן פונקציית ההתפלגות ניתן לחשב מדדים שונים של זמן ההישרדות, למשל תוחלת או חציון הזמן עד שקורה האירוע הבא, תוחלת הזמן בין שני אירועים עוקבים ועוד. המקור של ניתוח ההישרדות הוא מתחום הרפואה ומדעי החיים, אבל לניתוח הישרדות יש שימושים נרחבים גם מעבר לתחומים אלו. בדרך כלל, אם כי לא תמיד, האירועים מבטאים סוג מסוים של "כישלון" (failure). לדוגמה, בתחום הרפואה, משך הזמן עד לתמותה (או החלמה) ממחלה. בבעיות אמינות, משך הזמן עד שרכיב מסוים, או המערכת כולה, מפסיקים לפעול. בתחום הביטוח, משך הזמן עד להגשת תביעה. ביישומים בשיווק, משך הזמן לרכישת מוצרים ברי-קיימא (כגון כלי רכב). בתחום התקשורת, משך הזמן שעובר עד שלקוח נוטש לחברה אחרת (churning), ועוד.

במאמר הזה הרחבנו את הנושא של ניתוח ההישרדות גם לתחום מדעי החברה והניהול עבור אירועים תלויי זמן, עם דגש על בעיות חיזוי. דוגמה טיפוסית היא מבצעים לשיווק רכבים חדשים או מוצרים ברי-קיימא אחרים. לבעיות אלה יש שלושה מאפיינים: (1) המשתנה התלוי, או התגובה, הוא משך הזמן עד שקורה האירוע הבא. (2) התצפיות הן קטומות במובן זה שאין לנו את מלוא האינפורמציה על כל התצפיות בתקופת הניסוי. (3) זמני ההישרדות של התצפיות תלויים במשתנים מסבירים. פונקציית המטרה בבעיות אלה היא ללמוד על התנהגות הלקוחות על סמך מדגם למידה כדי לחזות את ההסתברות לאירועים עתידיים, מעבר לתקופת הניסוי, עבור לקוחות חדשים של לקחו חלק בבנייה של המודל.

הבעיה העיקרית היא שהגישות המסורתיות של ניתוח הישרדות מוגבלות רק לחיזוי בתקופת הניסוי ואינן בנויות לחזות אירועים מעבר לתקופת הניסוי, לא כל שכן אירועים בדידים כמו קניות של רכבים. על מנת להתמודד עם בעיות חיזוי מסוג זה, אנו משתמשים בגישה המשלבת בין מודלים של הישרדות ומודלים של חיזוי אנליטי מבוססי רגרסיה. הרעיון הוא להשתמש במודלים מהתחום של ניתוח הישרדות על מנת לאמוד את הפרמטרים של המודל הסטטיסטי, ואז להשתמש בעקרונות של גישות לחיזוי אנליטי על מנת לחזות אירועים עתידיים מעבר לתקופת הניסוי. הדגמנו את הגישה הזו על תחום הרכב על סמך קובץ נתונים סינתטי שיצרנו באמצעות סימולציה.

למיטב ידיעתנו, לא נעשה עדיין שימוש נרחב בגישות של ניתוח הישרדות בבעיות חיזוי בתחום הניהול ומדעי החברה עבור אירועים תלויי זמן, וגם הספרות המקצועית בנושא הזה

די דלילה. עד עכשיו יישמנו את הגישה המשולבת בתחום הרכב והביטוח. בתחום הרכב מדובר על פרויקט לאמידת ערך הלקוח (Lifetime value) של לקוחות חברת Ford במדינה מובילה באירופה, שבו השתמשנו בניתוח הישרדות על מנת לאמוד את משך הזמן עד לרכישת הרכב הבא עבור כל הלקוחות שבתחילת תקופת הניסוי היה ברשותם רכב של חברת Ford (Levin and Zahavi, 2006). בתחום הביטוח מדובר על בניית מערכת לקביעת תעריפי ביטוח עבור כלי רכב שמשקפים את הסיכון של בעל הרכב להיות מעורב בתאונה. הפרמטר שאמדנו באמצעות המודל לניתוח הישרדות הוא משך הזמן הממוצע בין שני תביעות עוקבות (MTBC – Mean Time Between Claims) עבור כל בעל רכב, ששימש כבסיס לקביעת תעריפי ביטוח מבוססי סיכון (Kahane et al., 2007).

מטרת המאמר הזה היא להציף את הפוטנציאל של שימוש במודלים של ניתוח הישרדות עבור אירועים תלויי זמן גם בתחומי הניהול ומדעי החברה, לעורר את המודעות לנושא הזה בתעשייה, וגם להציע פתרונות לשילוב מודלים של הישרדות בתהליכים של חיזוי אנליטי. יש לצפות שהצורך בכלים מתקדמים יותר לטיפול בבעיות תלויי זמן בניהול ומדעי החברה, בשילוב עם הגידול של כוח המחשוב וההתקדמות המדהימה של תחום הבינה המלאכותית והחיזוי האנליטי, ירחיבו בעתיד הקרוב את השימוש בכלים של ניתוח הישרדות גם בתחום הניהול ומדעי החברה.

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

רשימת מקורות

Cox D. R. (1972), Regression Models and Life Tables, *Journal of the Royal Statistical Society, Section B, 34*, pp. 187-220.

Emmert-Streib, F., and Dehmer, M. (2019). Introduction to survival analysis in practice. *Machine Learning and Knowledge Extraction*, *1*(3), pp. 1013-1038.

Gao, Y., Duan, W., & Rui, H. (2022). Does social media accelerate product recalls? evidence from the pharmaceutical industry. *Information Systems Research*, *33*(3), pp. 954-977.

Garg, S., Sinha, S., Kar, A. K. and Mani, M. (2022). A review of machine learning applications in human resource management. *International Journal of Productivity and Performance Management*, *71*(5), pp. 1590-1610.

Huayu, Li, Yong, G.C., Hengshu, Zhu, Hui Xiong and Hongke Zhao (2017), Prospecting the Career Development of Talents: A Survival Anaysis Perspective, *KDD 2017 research Paper*.

Kahane, Y., Levin, N., Meiri, R. and Zahavi, J. (2003), Applying Data Mining Technology for Insurance Rate Making: An Example of Automobile Insurance, *Asia-Paciifc Journal of Risk and Insurance, Vol. 2, Issue 1*, pp. 33-50.

Kaplan E.L. and Meire, P. (1958), Nonparametric Estimation from Incomplete Observations, *Journal of the American Statistical Association, 53*, pp. 457-481.

Kleinbaum, D. G. and Klein.M. (2011) Survival Analysis – A Self Learning Text, *Springer Publishers,* NY.

Levin, N. and Zahavi, J. (2005), GainSmarts – Data Mining System for Marketing, *Handbook of Data Mining and Knowledge Discovery,* Springer, edited by O. Maimon and L. Rokach, pp. 1261- 1301,.

Levin, N. and Zahavi, J. (2003), Ford Spain – LTV Model, *Urban Science/Q-Ware internal report,* Detroit, MI.

Miller R. G. (1997), *Survival analysis*, John Wiley & Sons.

Wang, Z., Jiang, C., and Zhao, H. (2022). Know Where to Invest: Platform Risk Evaluation in Online Lending. *Information Systems Research*, *33*(3), pp. 765-783.

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