הפקולטה לניהול ע"ש קולר
אוניברסיטת תל אביב
מאמר זה פורסם בכתב העת "חידושים בניהול", גיליון מס' 7, אוקטובר 2020,
היוצא לאור על ידי הפקולטה לניהול על שם קולר באוניברסיטת תל אביב.

לחצות את הגשר – מביג דאטה לתובנה עסקית

יעקב זהבי

רונית הנגבי

ד"ר רונית הנגבי היא בעלת דוקטורט בחינוך מאוניברסיטת תל אביב, מתמחה במודלים של משוואות מבניות (SEM) ושילובם בתחום נתוני העתק. עם סיום לימודיה התמחתה במכון מקס פלאנק בברלין ובהמשך פנתה לתעשיית ההיי-טק. ד"ר הנגבי היא מחלוצות ה-Web Mining ופיתחה יישומים להנגשת נתונים למקבלי החלטות ואנליסטים. בשנת 1999 הקימה את חברת Eastat בישראל ובשנת 2006 הקימה את חברת WerWater בהולנד, ובמסגרתן פיתחה מתודולוגיות לפילוח וטרגוט לקוחות. על לקוחותיה נמנות חברות בין-לאומיות בתחום הפיננסים, הטלקום, הרכב והמסחר מהגדולות בעולם. כיום היא יועצת לחברות Big Data ולאחרונה השיקה מיזם בתחום ה-Food Tech העוסק באופטימיזציה של תזונה קיטוגנית.

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 כמנתח 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 "הסבה" לתחום של שיווק מבסיסי נתונים, וכך הגיע לתחום של כריית מידע שבו הוא עוסק עד היום. זכה פעמיים ברציפות במדליית הזהב בתחרות השנתית לגילוי ידע שמאורגנת על ידי ה-American Machinery Computation. מספר מאמרים שלו בתחום זה זכו בפרסים על מצוינות אקדמית.

תקציר

הממדים והסיבוכיות של בניית מודלים לחיזוי אנליטי בעולם נתוני העתק (Big Data), הביאו בשנים האחרונות לתמורה בבניית מודלים של חיזוי בצורה אוטומטית: פחות התערבות אנושית, ויותר גישות המבוססות על למידת מכונה. המטרה היא להנגיש את הטכנולוגיה הזו לכלל מקבלי ההחלטות גם בארגונים עסקיים שאין להם יכולת להתמודד עם הסיבוכיות של בניית מודלים רב-ממדיים לחיזוי. אלא שתהליך זה מניב בדרך כלל מודלים "סגורים" שאינם מאפשרים לראות מהם המשתנים והמאפיינים שהשפיעו על תהליך החיזוי וללמוד מתוך המודל. זאת בניגוד למודלים "פתוחים" המאפשרים, באמצעות ניתוח לעומק של מרכיבי המודל, לקבל יותר תובנות ולהגדיל את מרחב ההחלטות גם לכיוונים עסקיים ושיווקיים נוספים.
במאמר זה נרחיב את הדיון על ההבדלים בין מודלים "פתוחים" למודלים "סגורים", ובאמצעות שלושה אירועים אמיתיים של חיזוי אנליטי בתחומים שונים, נדגים כיצד ניתן לנצל תובנות הנובעות ממודלים פתוחים של חיזוי אנליטי כדי לקדם החלטות עסקיות ומהלכים שיווקיים ובכך לגשר בין עולם נתוני העתק לבין העולם העסקי.

הקדמה

אנחנו מצויים היום בעיצומו של עידן נתוני העתק המתאפיין בהררים של נתונים שמגיעים מכל עבר – רשתות חברתיות, בלוגים ופורומים באינטרנט, IoT, חברות עסקיות, רשויות ממשלתיות, עיתונים מקוונים ולא מקוונים, חיישנים ומצלמות מסוגים שונים, וזו רק רשימה חלקית. אפילו הטלפונים החכמים שאנחנו נושאים בכיס משדרים מידע בלי הרף, גם כשהם כבויים. כמות עצומה זו של נתונים הביאה לאחרונה לפיתוח מואץ של מגוון רחב של מודלים אנליטיים במטרה לנתח ולהפיק מידע מהררי הנתונים האלה. במאמר זה נשים את הדגש על מודלים של חיזוי אנליטי (PA – Predictive Analytics), העוסקים בחיזוי שיעורי התגובה של אירועים עתידיים על סמך תצפיות מהעבר שעבורן ערכי התגובה ידועים. תהליך החיזוי מורכב למעשה משני תהליכים עיקריים: תהליך של בניית מודל אנליטי על מנת להתאים מודל לנתונים, ותהליך של ציינון (Scoring) – חיזוי התגובה של צרכנים חדשים על בסיס המודל האנליטי.

אם בעבר תהליך בניית המודל בתחום של כריית מידע התבסס על שילוב של Art & Science, הרי שהיום פחתה ההתערבות האנושית בבניית מודלים לטובת גישות המבוססות על למידת מכונה (ML – Machine Learning) שלהן יכולת עיבוד נתונים מהירה ומדויקת יותר, על מנת לבנות מודלים של חיזוי בצורה אוטומטית. הפלט של מודלים אלה הוא לרוב משוואות או חוקים שלוקחים חלק בתהליך קבלת ההחלטות. המטרה היא להנגיש את הטכנולוגיה הזו לכלל מקבלי ההחלטות גם בארגונים עסקיים בינוניים וקטנים שאין להם את היכולת והמשאבים להתמודד עם בנית מודלים רב-ממדיים לחיזוי.

לכאורה, התהליך הזה נראה כמו החלום הוורוד של כל משתמש עסקי, שכן הוא מאפשר לו לקבל מודלים איכותיים של חיזוי במהירות וביעילות. ואכן, המהירות והדיוק של גישות למידת מכונה הן ללא ספק יתרונות בולטים של השינויים שחלו בתחום בשנים האחרונות. אלא שתהליך זה הוא בבחינת "קופסה שחורה" מפני שהוא מתעלם מהתרומה החבויה במרכיבים של המודל שיכולים להניב משמעות עסקית חשובה. המודלים המתקבלים הם מודלים "סגורים" שאינם מאפשרים לראות מהם המשתנים והמאפיינים שהשפיעו על תהליך החיזוי וללמוד מתוך המודל. מדובר כאן על תהליך טכני בעיקרו, שמשתמש במשוואות ובחוקים שנוצרו על ידי תהליך למידת מכונה על מנת לחשב את הציונים שעליהם מתבססות ההחלטות העסקיות מבלי לרדת למהות של המודל ולמאפיינים שלו. לעיתים מזומנות

כל התהליך הוא שקוף למשתמש, וכל מה שהמשתמש "רואה" אלה רק הציונים שמקבלים הלקוחות החדשים, למשל הציונים של הלקוחות במבצעי שיווק.

לעומתם, מודלים פתוחים משלבים Art & Science ומבוססים גם על למידת מכונה, שזה מרכיב ה-Science של התהליך, וגם על ידע אנושי הנוגע לתחום של הבעיה (Domain knowledge), שמכניס את ממד ה-Art לתהליך. מודלים פתוחים גם מאפשרים הצצה אל תוך ה"קרביים" של המודל, שכן כל מרכיבי המודל חשופים למשתמש. למשל, במודל של רגרסיה (לינארית או לוגיסטית), שהוא ללא ספק מודל החיזוי הוותיק והנפוץ ביותר, המשתמש רואה את משוואות הרגרסיה, את האומדים של מקדמי הרגרסיה והמובהקות שלהם, את רמת הדיוק של המודל, את מקדם הקביעה (R-Square), וכן מדדים הנובעים מהיישום של המודל לצורך קבלת החלטות – תרשימי רווחים וטבלאות רווחים (Gains charts/Gains tables), מדדים להערכת התאמת יתר (Over fitting), עקומות ROC ועוד. יש לכך חשיבות ראשונה במעלה בתהליך קבלת ההחלטות, מכיוון שבאמצעות ניתוח לעומק של תוצאות המודל אפשר להגדיל את מרחב ההחלטות לכיוונים עסקיים נוספים, ולספק מענה לסוגיות עסקיות "מפתיעות" שלרוב אינן אינטואיטיביות ולעיתים גם לא צפויות (מה שמכונה Quick wins) ובכך לגשר בין נתוני העתק לבין בעיות ותהליכים עסקיים.

במאמר זה נרחיב את הדיון על השימוש במודלים פתוחים כדי לענות על מגוון של שאלות עסקיות, ונראה איך באמצעות "מבט מעבר למודל" וניתוח לעומק של תוצאות המודל אפשר לקבל תובנות נוספות שיש להן משמעויות עסקיות חשובות. תנאי הכרחי לחשיבה מחוץ למודל הוא בניית מאגרי נתונים (Data lakes) שמשלבים נתונים ממקורות שונים, לאו דווקא כאלה שנדרשים עבור מודל החיזוי. משימה זו דורשת הכרה עמוקה של תחום הבעיה, של מודל החיזוי ושל בסיסי הנתונים הפנימיים של הארגון, וכן נתונים חיצוניים רלוונטיים. לכן היא מחייבת מעורבות פעילה של מדעני הנתונים (Data scientists) בכל התהליך, בשילוב עם המשתמשים העסקיים, כבר משלב אפיון הבעיה והגדרת הנתונים הרלוונטיים הקשורים אליה. שיתוף פעולה כזה יכול להניב תועלת רבה להבנת המודל ולשימוש בתוצאותיו לצורך קבלת החלטות עסקיות ושיווקיות.

לצורך הבהרת הנושא נתמקד במודלים של חיזוי אנליטי בעולם השיווק, ונדגים את הנושא באמצעות מספר אירועים אמיתיים (Use cases) מתחום תעשיית הרכב, בנקאות השקעות ותקשורת סלולרית. היריעה במאמר זה קצרה מלדון באספקטים התיאורטיים והמעשיים של פיתוח ויישום מודלים של חיזוי אנליטי לקבלת החלטות. דיון מפורט של תהליך החיזוי האנליטי מופיע אצל זהבי (2017).

במאמר הזה אנו מתרכזים במודלים פתוחים, אך אי אפשר להתעלם מהנושא של מודלים סגורים. מתברר שהסיבוך של גישות בינה מלאכותית (AI) ולמידת מכונה, שרובן ככולן מניבות מודלים סגורים, עשוי להרתיע את המשתמשים העסקיים ומקבלי ההחלטות מליישם את המודלים האלה במציאות, מפני שהמודלים המתקבלים הם לחלוטין לא ברי הבנה ולא אינטואיטיביים. מציאות זו "הולידה" לאחרונה כיוון מחקרי חדש שנקרא אינטליגנציה מלאכותית ברת-הסבר (Explainable AI), הידועה גם בראשי תיבות XAI. המטרה היא לא רק להסביר מודלים של בינה מלאכותית ולמידת מכונה, אלא גם להסביר את כל התהליך כולו, כולל הנתונים שהביאו לבנית המודל, על מנת להעלות את האמון של מקבלי ההחלטות במודלים השונים ולעודד אותם ליישמם בפועל.

קימות שתי גישות עיקריות ל-XAI: ניתוח מראש (Ante-hoc) וניתוח בדיעבד (Post-hoc). גישות הניתוח מראש מכניסות מנגנונים להסבר המודלים כחלק אינטגרלי של בניית המודל. גישות הניתוח בדיעבד מנסות להסביר ולאשש את המודלים לאחר בניית המודל. גישות XAI נידונות בהרחבה בספרות, וסקירה טובה של הנושא ניתן למצוא במאמר המקוון של Gandhi (2019) ובמאמר המקוון של Schoenborn and Althoff (2019). מתוך שלל הגישות שהוצעו על מנת "לפרש" מודלים סגורים, אנו נדון בשני מדדים להערכת משתנים מסבירים במודלים סגורים של חיזוי אנליטי – מדד המבוסס על גישת LIME, ומדד המבוסס על ערך שפלי (Shapley value).

לבסוף נעיר שכל הטכניקות שנציג להלן להסברת מודלים פתוחים וסגורים שייכים לקטגוריה של Post-hoc (ניתוח בדיעבד), שכן כל הניתוח נעשה בדיעבד, לאחר בניית מודל החיזוי.

מעבר למודל – ניתוח המידע הצפון במודלים פתוחים לקבלת החלטות עסקיות

רוב המודלים בחיזוי אנליטי מכוונים לשיווק פרסונלי, המכונה בידי אנשי השיווק 1:1 (One-to-one) או "מתחת לקו" (BTL– Below The Line), וזאת להבדיל ממודלים "מעל לקו" (ATL– Above The Line) המיועדים לבניית מהלכי פרסום המוניים, מיתוג והחלטות אסטרטגיות. בסעיף זה נתאר כיצד ניתן לממש את הפוטנציאל הגלום בשימוש במערכות פתוחות במודלים מבוססי רגרסיה למהלכים טקטיים של איתור קהלי יעד לשיווק של מוצרים ושירותים ("טרגוט"), וגם למהלכים אסטרטגיים ולתכנון מבצעי פרסום.

בסעיף זה נדון בכמה היבטים:

- תכלול מודלים
- מדרוג המשתנים במודל
- המקדמים ורמות המובהקות של המשתנים
- סיבתיות (Causation)
- טרנספורמציות ואינטראקציות
- ניתוח פרטני
- אמינות המודל

תכלול מודלים

גישת התכלול (Ensemble) (Rocach, 2010) היא גישה מקובלת בכריית מידע ומטרתה להקטין את טעות החיזוי, בדומה לגישה בסטטיסטיקה של ניתוח שאריות (Residual analysis), ועל ידי כך להפיק תחזיות מדויקות יותר. בשלב הראשון מריצים מספר מודלים שונים במקביל על אותו קובץ אימון ומחשבים את הציון של כל תצפית בבסיס הנתונים עבור כל מודל בנפרד. בשלב השני מחשבים את התחזית לכל תצפית באמצעות ממוצע הציונים של המודלים השונים.

גישת התכלול הנפוצה ביותר היא גישת ה-Random forest (Hastie et al, 2009), שלפיה התחזית עבור כל תצפית מתקבלת על סמך ממוצע הציונים של כמה עשרות או מאות מודלים של עצי החלטה שמריצים על מדגמים אקראיים שונים מתוך קובץ האימון (Training dataset). ההנחה שעומדת בבסיס גישת התכלול היא שאם משהו "מתפספס" במודל אחד, הוא יבוא לידי ביטוי במודל אחר, ולכן תחזית המתבססת על מיצוע הציונים של כמה מודלים תיתן תחזית מדויקת יותר.

גישת תכלול אחרת שזוכה לאחרונה לפופולריות היא גישת GB (Gradient Boosting) שמורכבת מסדרה של מודלים שמריצים בזה אחר זה (Friedman, 1999). בשלב הראשון בונים מודל (בדרך כלל עצי החלטה או רגרסיה לוגיסטית) עם כל המשתנים המקוריים, ומחשבים עבור כל תצפית את השארית בין הערך האמיתי של המשתנה התלוי והערך החזוי שלו. בשלב השני מריצים מודל חיזוי נוסף שבו מסבירים את השאריות שחושבו במודל הקודם באמצעות המשתנים המסבירים, וכך הלאה.

אפשרות נוספת היא "לדלות" את הלקוחות שלא נכללו במבצע השיווק על ידי המודל הראשון באמצעות חקירה לעומק של המאפיינים שלהם, על מנת להפיק מידע שיווקי מועיל. לדוגמה, ניתוח לעומק של הפעילות הסלולרית בשעות 20:00-18:00 באירוע תקשורת סלולרית (שיתואר בהרחבה בהמשך), הצביע על סגמנטים מיוחדים באוכלוסייה בעלת מאפיינים שיווקיים ייחודיים. דוגמה אחרת היא חברת רכב שהשתמשה במשתנה המיקום הגיאוגרפי, שלא בהכרח נדרש עבור מודל החיזוי, על מנת להימנע מלהציע מכוניות חשמליות ללקוחות באזורים שאין בהם עמודי טעינה זמינים.

ואומנם, גישות תכלול מחייבות מבט "מעבר למודל" בכך שהן מצביעות על מאפיינים ותכונות שלא נצפו על ידי המודלים המקוריים, ובאמצעות שילובם במודלים עוקבים או ניתוח עומק שלהם, הם מאפשרים לקבל תובנות עסקיות נוספות ואף להרחיב את מעגל הלקוחות במבצעי השיווק.

מדרוג המשתנים במודל

מדרוג המשתנים שנכנסו למודל הרגרסיה מעיד על מידת החשיבות וההשפעה של המשתנים שלהם במודל. בדרך כלל המשתנים מדורגים בסדר יורד, מהמשתנה המשפיע ביותר

למשתנה הפחות משפיע. את החשיבות של המשתנה מודדים באמצעות ההשפעה שלו על מדד טיב ההתאמה שבו השתמשנו על מנת לבנות את המודל. למשל, ממוצע שורש השגיאות הריבועיות (RMSE – Root Mean Square Errors), מקדם הקביעה R2, פונקציית הנראות (Likelihood function) או הלוגריתם הטבעי שלו, ואחרים. טבלה 1 מציגה את דירוג המשתנים ממודל "טרגוט" של מועדון ספרים שמציע ללקוחותיו לרכוש ספר חדש בשם "ההיסטוריה של האומנות בפירנצה". המדד לחשיבות המשתנים בתרשים זה הוא הלוג של פונקציית הנראות. מטבלה זה עולה שהמשתנה המשפיע ביותר הוא n.art המבטא את מספר ספרי האומנות שהצרכן רכש בעבר. הסרת המשתנה הזה מהמודל, כאשר כל יתר המשתנים והפרמטרים נשארים זהים, מקטינה את הלוג של פונקציית הנראות ב-31.95%. נעיר שפונקציית המטרה בבניית מודל רגרסיה היא למקסם את פונקציית הנראות ולכן ירידה בערך של פונקציית הנראות בכמעט 32% נותנת מודל גרוע יותר עם תחזיות שגויות יותר. להבנת המשמעות של המשתנה הזה חשוב לדעת מה היה הסימן של המשתנה הזה במודל הרגרסיה. במקרה של המשתנה n.art המקדם של המשתנה הוא חיובי – והמשמעות היא שככל שהצרכן קנה בעבר יותר ספרי אומנות, כך גדלה ההסתברות שלו להיענות להצעה השיווקית הנוכחית, כלומר לרכוש את הספר "ההיסטוריה של האומנות של פירנצה".

המשתנה הבא בתור הוא Recency – משך הזמן שעבר מאז הקנייה האחרונה (בחודשים) של הצרכן. הסרת המשתנה הזה מהמודל, כאשר כל יתר המשתנים והפרמטרים נשארים זהים, מקטינה את הלוג של פונקציית הנראות ב-23.21%. כאן המקדם של המשתנה שלילי, כלומר ככל שעבר זמן רב יותר מאז הקנייה האחרונה של הלקוח, ההסתברות שהוא ייענה להצעה השיווקית לקנות את הספר הנוכחי יורדת. וכך הלאה לגבי יתר המשתנים.

מבחינה עסקית, מדרוג המשתנים מאפשר לאתר הזדמנויות שיווקיות נוספות. למשל, בדוגמה שלנו להשיג רשימות של אנשים שקנו בעבר ספרי אומנות ולהציע להם ספרי אומנות נוספים. ואם מדובר על משתנים בלתי תלויים שיכולים גם לשמש כמשתני החלטה (כגון תקציב השיווק), הרי נרצה לתת עדיפות למשתנים המשפיעים ביותר שנמצאים בראש טבלת הדירוג.

1 טבלה 1 נלקחה מתוך חבילת התוכנה לחיזוי אנליטי של חברת DMWay Analytics.

טבלה 1: מדרוג המשתנים עבור מועדון הספרים על פי מודל החיזוי

Variable	% explanation of log likelihood
n.Art	31.95%
Recency	23.21%
n.History	16.00%
Gender	11.75%
Freqency	4.29%
n.Reference	2.56%
Purchese.Secret.of.italian.k	0.82%
Money	0.72%
n.Cooking	0.56%
n.DIY	0.37%
Purchese.Italian.Art	0.22%
Purchese.Atlas.of.Italy	0.20%
n.Teen	0.12%

מקדמים ורמות המובהקות של המשתנים

כפי שצוין לעיל, המקדמים של המשתנים במודל רגרסיה מעידים על תרומתם למודל. משמעות התרומה של משתנה שונה ממודל למודל. למשל, במודל רגרסיה ליניארית המקדם של המודל מבטא את השינוי בערך של המשתנה התלוי כתוצאה משינוי של יחידה אחת בערך של המשתנה הבלתי תלוי. המקדם של משתנה במודל יכול להיות חיובי או שלילי. מקדם חיובי אומר ששינוי ביחידה אחת בערך של המשתנה הבלתי תלוי מעלה את הערך של המשתנה התלוי, ולהיפך במקרה של מקדם שלילי.

רמת המובהקות הכללית, שבדרך כלל מוגדרת מראש (לרוב 5% ולעיתים גם 1%), קובעת את רמת הסף לבדיקת המובהקות של המשתנים במודל. רמת המובהקות של משתנה נקבעת על סמך רמת המובהקות האפקטיבית שלו הידועה בשם p-value. ככל שה- p-value של המשתנה קטן יותר, והוא גם נמוך מרמת המובהקות הכללית, המשתנה תורם יותר למודל, כלומר הוא "מובהק" יותר.

העובדה שבניית מודל מבוססת על מדגם גורמת לשני סוגי טעויות: טעות מסוג ראשון – להכניס למודל משתנים שאינם מובהקים. טעות מסוג שני – הוצאה של משתנים מובהקים מהמודל. מאחר שאי אפשר להקטין בו-בזמן גם את הטעות מהסוג הראשון וגם את הטעות מהסוג השני, יש צורך בשקלול תמורות (Tradeoff analysis) על מנת למצוא את התמהיל האופטימלי של משתנים שיש להכניס למודל. בדרך כלל אין ספק שיש להכניס למודל משתנים מובהקים במיוחד, כלומר משתנים עם רמת p-value נמוכה מאוד וקרובה לאפס. הבעיה העיקרית היא עם משתנים "גבוליים", כלומר משתנים עם רמת p-value הנעים סביב רמת המובהקות הכללית. לרוב מקובל להוציא משתנים כאלה מהמודל, אבל מתברר שגם למשתנים אלה יכולה להיות תרומה משמעותית (במיוחד אם יש הרבה מהם). במסגרת האירועים שבהם נדון בהמשך נביא דוגמאות כיצד ניתן להשתמש במשתנים כאלה לצורך החלטות עסקיות.

ראוי לציין שמשתנים מובהקים שנכנסים למודל אכן יכולים להיות אלו שמסבירים את התופעה הנחקרת, או ששינוי בערכם (אם יש לנו שליטה עליהם) יכולים להשפיע על התוצאה. מצד שני, משתנים יכולים להיכנס לתוך המודל בתור משתנים מובהקים רק כי הם מתואמים עם הגורם המשפיע האמיתי (שעליו לא תמיד יש לנו שליטה), כתוצאה מהטיות שנובעות ממשתנים חשובים שנשמטו מהמודל (או מסיבות אחרות). זוהי נקודה עדינה שמשמעותה, מבחינה מעשית, היא שמשתנה שיצא מובהק לא בהכרח מסביר את התופעה אלא מצביע על כך שיש לחקור את המשתנה הזה יותר לעומק כדי לוודא שהוא אכן נכנס למודל בדין ויש לו קשר מובהק למשתנה התוצאה (המשתנה התלוי).

בהקשר הזה ראוי לציין את המחקרים של גלית שמואלי בנוגע להבדל בין מודל הסבר למודל חיזוי (Shmueli, 2010; Shmueli and Koppious, 2011), שמהם משתמע שמודל הסבר הוא לא בהכרח מודל חיזוי טוב. למסקנה זו יש השלכה על רמות המובהקות שיש להגדיר עבור מודל החיזוי, שהן בדרך כלל מקלות יותר מאשר במודל הסבר כדי למנוע מצב של התאמת יתר. למשל, רמת מובהקות של 10%, שהיא לא מקובלת עבור מודל הסבר, יכולה להיות רלוונטית למודל חיזוי. דיון נרחב בהבדלים בין מודל הסבר למודל חיזוי והמשמעויות שלהם מופיע אצל זהבי (2017).

סיבתיות (Causation)

בשל החשיבות של קשרים סיבתיים בקבלת החלטות, נושא הסיבתיות (Causation) תפס תאוצה מרשימה בשנים האחרונות. הכהן הגדול של חקר הסיבתיות הוא המדען יהודה פרל שפיתח "שפה" מיוחדת (Structural Causal Modeling – SCM) לניתוח קשר סיבתי בין משתנים המתבססת על תורת הגרפים ורשתות הסתברותיות (Pearl, 2000).

עולם נתוני העתק פיתח לאחרונה כיוונים נוספים לאפיון סיבתיות בין משתנים באמצעות תכנון ניסויים, במטרה לאמוד את ההשפעות של טיפולים על משתנה התוצאה, וכן להתמודד עם העובדה שטיפולים שונים משפיעים בצורה שונה על קבוצות אוכלוסייה שונות (Athey and Guido, 2016; Wagner and Athey, 2018).

בהקשר שלנו נציין שקשרים סיבתיים בין משתנים לא באים לידי ביטוי במודלים של רגרסיה וגישות ML אחרות, שלרוב מתבססים על קשרים סטטיסטים וקורלציות בין משתנים. כך למשל, מתאם גבוה בין משתנה תלוי ובין משתנה בלתי תלוי, שקרוב לוודאי ייכנס למודל הרגרסיה בתור משתנה מובהק, לא בהכרח מעיד על קשר סיבתי בין המשתנים. מאידך, קשר סיבתי קיים רק בין שני משתנים שמתואמים ביניהם. לכן יש צורך להיזהר ממסקנות על קשרים סיבתיים בין משתנים במודל החיזוי המבוססות על קורלציות ורמות מובהקות של המשתנים. כאן נכנס לתמונה מדען הנתונים שנדרש להביט "מעבר למודל" החיזוי כדי לאפיין, תוך שילוב של תחום הבעיה, ובעצה אחת עם המשתמשים העסקיים, את המשתנים במודל שקיים ביניהם לא רק קשר סטטיסטי מובהק אלא גם קשר סיבתי. כמו כן, היכרות טובה עם המודל ועם המשתנים המסבירים בבסיס הנתונים יכולה לעזור באיתור המשתנים המערבלים (Confounder variables) ובכך להימנע ממסקנות מוטעות הנובעות מהשמטה מהמודל של המשתנים האלה.

טרנספורמציות ואינטראקציות

ההנחה הסמויה במודל חיזוי לינארי, כגון רגרסיה ליניארית או לוגיסטית, היא שהקשר בין המשתנה התלוי לבין המשתנה

הבלתי תלוי הוא ליניארי. עם זאת, במציאות עצמה הקשר בין המשתנים הוא לעיתים לא ליניארי – לדוגמה הקשר בין רמת ההכנסה (משתנה תלוי) לבין גיל הלקוח (משתנה בלתי תלוי). מניסיון אמפירי ידוע שבתחילה רמת ההכנסה עולה ככל שהגיל עולה עד שהיא מגיעה לרמה מרבית, ואז היא הולכת ויורדת עם הגיל. לקשר זה צורה של פרבולה, הרחוקה מאוד מלהיות ליניארית. במצבים כאלה מטפלים בדרך כלל באמצעות טרנספורמציות לא ליניאריות של המשתנים במודל הרגרסיה. למשל, בדוגמה הנ"ל מגדירים את משתנה רמת ההכנסה כפונקציה פרבולית של משתנה הגיל.

אינטראקציות בין משתנים בלתי תלויים הן סוג אחר של טרנספורמציות. בדרך כלל אינטראקציה במודל רגרסיה היא מסדר שני שבה מעורבים שני משתנים בלתי תלויים. אבל לא מן הנמנע להגדיר אינטראקציות מסדר שלישי שבה מעורבים שלושה משתנים בלתי תלויים. דוגמה לאינטראקציה מסדר שני היא אינטראקציה בין משתנה קטגורי כגון המגדר – זכר או נקבה, ומשתנה רציף כגון רמת השכר. אינטראקציה בין המשתנים האלה קיימת כאשר השיפוע של רמת השכר (כלומר הגידול בערך של המשתנה התלוי עבור גידול של יחידה אחת ברמת השכר) הוא שונה בין גברים לנשים. אינטראקציה בין משתנים מיוצגת במודלים של רגרסיה באמצעות המכפלה של המשתנים המעורבים, ובדוגמה של מגדר ורמת שכר באמצעות המכפלה female*salary, כאשר "female" הוא משתנה קטגורי שמקבל את הערך 1 עבור אישה ואת הערך 0 עבור גבר, ואילו "salary" מבטא את השכר של הלקוח.

דווקא משום שמספר האינטראקציות בעולם נתוני העתק יכול להגיע לעשרות אלפים ויותר, הבנת האינטראקציות מאפשרת למשתמש להעשיר את התובנה העסקית. לא מעט מהממצאים המפתיעים ביותר בפרויקטים נמצאים דווקא באינטראקציות, כי הן מרמזות על קלסיפיקציות ייחודיות. עובדה זו מדגישה ביתר שאת את המידע החבוי בנתונים ויכולה להצביע על הזדמנויות עסקיות נוספות מהסוג של Quick wins.

ניתוח פרטני

עד עכשיו התייחסנו לשאלות הנוגעות לכלל המודל או למשתנים בתוכו. בסעיף זה נדון בניתוח פרטני של תוצאות המודל שמאפשרות לרדת לרמה של הלקוח הבודד במטרה להבין יותר את ההתנהגות שלו, ואם אפשר – גם להשפיע על התנהגותו.

בטבלה 2 שלפנו קטע להדגמה מתוך תוכנת החיזוי עבור מועדון ספרים. כל שורה בטבלה זו מתייחסת ללקוח ספציפי מסוים. כל עמודה מציינת את התרומה בסולם אורדינלי של משתנה ספציפי עבור הלקוח. עמודת ה-account מציינת את מספר הלקוח, עמודת ה-score את ה"ציון" של התצפית ממודל הרגרסיה (למשל, ברגרסיה לוגיסטית, הסתברות הרכישה של הלקוח). עמודת ה-gender את התרומה של משתנה המגדר על הציון, עמודת ה-money את התרומה של משתנה ההוצאה הכספית של הלקוח על רכישות שביצע בעבר במועדון הספרים על הציון, וכך הלאה. נדגיש שמדובר כאן על מדרג של התרומה של המשתנים המסבירים עבור כל תצפית (לקוח), ולא על תרומות מוחלטות. טבלה זו מכילה רק את המשתנים המסבירים המקוריים שנכנסו למודל הרגרסיה. ככל שהערך של התרומה (בערך מוחלט) גבוה יותר, ההשפעה של המשתנה הרלוונטי גבוהה יותר. סימן חיובי מתאר השפעה חיובית של המשתנה המסביר עבור הלקוח הספציפי על הציון של הלקוח, וההיפך לגבי ערכים שליליים. לנוחיות הקריאה של הטבלה ניתן להוסיף לה צבעים, בדומה לפונקציית ה-Condition שקיימת ב-Excel. למשל, צבע אדום – עבור תרומות שליליות גבוהות, וצבע ירוק – עבור תרומות חיוביות גבוהות, וכך הלאה, ולקבל טבלה מהסוג של Heat Map שמקלה על הקורא לקרוא את הטבלה.

כאמור, טבלה זו מאפשרת ניתוח לעומק ברמה של הלקוח הבודד. לדוגמה, ביישומים פיננסיים, אם לקוח ספציפי מפגר בתשלומי הלוואות או שהפסיק לשלם אותן, לנסות להבין מדוע זה קורה ומהם המשתנים שמסבירים את התופעה הזו. יישום אפשרי נוסף הוא לצורכי שיווק, למשל לבחור מתוך רשימת הלקוחות רשימה של לקוחות עם תרומה חיובית גבוהה של משתנה מסוים ולהציע להם מוצרים משלימים, מה שמכונה מכירה צולבת (Cross-sell).

2 טבלה 2 נלקחה מתוך חבילת התוכנה לחיזוי אנליטי של חברת DMWay Analytics

טבלה 2: Heat Map

Account	Score	Gender	Money	Recency	Frequency	Age	N.Children	N.Cooking	N.DIY	N.Reference	N.Art	N.History
10003	0.013	-0.309	0.017	-0.264	0.285	-0.024	0.026	-0.229	0.136	-0.089	-0.483	-0.390
10005	0.015	-0.309	-0.045	-0.264	0.422	-0.115	0.026	-0.035	0.136	-0.089	-0.483	-0.390
10006	0.038	-0.309	-0.142	0.539	0.422	-0.115	0.026	0.197	0.136	-0.089	-0.483	-0.390
10008	0.059	-0.309	0.036	1.141	-0.941	0.112	0.026	-0.229	-0.943	-0.089	-0.483	1.763
10009	0.346	-0.309	0.249	0.740	-0.941	0.112	0.026	-0.397	-0.529	-0.089	3.201	0.327
10012	0.028	-0.309	0.020	-0.063	0.285	-0.115	0.026	0.197	0.136	0.194	-0.483	-0.390
10014	0.113	0.717	0.036	0.740	0.116	-0.070	0.026	-0.229	0.136	0.194	-0.483	-0.390
10015	0.017	0.717	-0.150	-1.267	0.422	0.112	0.026	-0.035	0.136	-0.089	-0.483	-0.390
10017	0.056	-0.309	0.028	0.539	0.116	-0.024	0.026	-0.229	0.136	-0.089	-0.483	0.327
10019	0.017	-0.309	0.036	0.138	0.285	-0.070	0.026	-0.035	-0.115	-0.089	-0.483	-0.390
10020	0.021	-0.309	0.041	0.138	0.285	-0.115	0.026	-0.035	0.136	-0.089	-0.483	-0.390
10021	0.069	0.717	0.036	0.338	0.285	-0.115	0.026	-0.035	0.136	-0.089	-0.483	-0.390
10022	0.023	-0.309	-0.150	-0.264	0.422	-0.115	0.026	0.197	0.136	0.194	-0.483	-0.390
10023	0.097	0.717	0.195	-1.066	-0.750	0.112	0.026	-0.397	-0.115	0.194	0.774	0.327
10024	0.312	-0.309	-0.139	1.141	0.285	-0.115	0.026	0.197	0.136	-0.089	0.774	0.327
10025	0.020	-0.309	0.036	-0.063	0.285	0.021	0.026	-0.035	0.136	-0.089	-0.483	-0.390
10026	0.017	-0.309	-0.150	-0.264	0.422	-0.115	0.026	0.197	0.136	-0.089	-0.483	-0.390
10027	0.186	0.717	0.084	0.138	-0.559	0.112	0.026	-0.229	-0.529	-0.089	0.774	0.327
10028	0.013	-0.309	0.036	-0.063	-1.036	0.112	0.026	-0.902	-0.115	0.194	-0.483	1.045

אמינות המודל

ולבסוף, אין ספק שהשיקול המרכזי שיקבע באיזו מידה מקבלי ההחלטות והמשתמשים העסקיים יתרגמו את תוצאות מודל החיזוי גם להחלטות עסקיות היא מידת האמון שלהם במודל. גישות אוטומטיות לבניית מודלים לחיזוי מכילות גם מנגנונים פנימיים לבדיקת איכות החיזוי של המודל. הדרך המומלצת היא לחלק באופן מקרי את הקובץ שמשמש לבניית המודל לשני תת-קבצים בלתי תלויים – קובץ אימון (Training dataset) וקובץ תיקוף (Validation dataset), לבנות את המודל על סמך קובץ האימון ולתקף אותו על סמך קובץ התיקוף (זהבי, 2017). אם ערכי החיזוי קרובים מספיק לערכים האמיתיים, ויש לכך גם מבחנים סטטיסטיים, אנו מסיקים שהמודל עובר את מבחן האמינות וניתן להשתמש בו לצורך חיזוי הערכים של המשתנה התלוי גם לגבי נתונים חדשים.

דרך נוספת לבדיקה של נכונות מודל החיזוי שתקפה למודלים פתוחים היא באמצעות בדיקות מהסוג של Face validity של הפרמטרים והמשתנים במודל החיזוי. בדיקות אלה כוללות, בין היתר, זיהוי משתנים לא רלוונטיים שנכנסו למודל (למשל מספר סידורי של התצפית), איתור וטיפול בתצפיות חריגות שעלולות לשבש את מודל החיזוי, סינון רעשים (למשל משתנים שמקבלים את אותו ערך עבור רוב התצפיות), משתנים מסבירים עם קורלציה גבוהה במיוחד עם המשתנה התלוי, סימנים לא סבירים של מקדמי הרגרסיה (למשל, משתנה מחיר עם מקדם חיובי שמנוגד לחוק הכלכלי שלפיו ככל שהמחיר עולה, הכמות המבוקשת קטנה), ועוד. רוב המבחנים האלה הם אינטואיטיביים, מובנים וקל להסביר אותם למקבלי ההחלטות. הניסיון האמפירי מוכיח שככל שהמודל מובן יותר למשתמשים העסקיים, כך גדל האמון שלהם במודל, וגדלה הנכונות שלהם להשתמש במודל בפועל לקידום הפעילות העסקית שלהם.

יישומים שיווקיים של מודלים פתוחים בעולם נתוני העתק

בסעיף זה נציג מספר אירועים אמיתיים המבהירים כיצד מבט מעבר למודל מביא לתובנות עסקיות נוספות שיכולות לעזור למנהלי השיווק לבדל מוצרים, למצוא פלחים באוכלוסייה עם מאפיינים שיווקיים מיוחדים, לאתר הזדמנויות ל-quick wins ולשפר את הרווחיות של הארגון.

מודלים של חיזוי לפרסום: תעשיית הרכב

האירוע הזה עוסק בחברת רכב בין-לאומית המייצרת רכבים לשימוש פרטי ומסחרי. על מנת לענות על כל צורכי ההנהלה, לרבות בניית אסטרטגיה שיווקית מבוססת נתונים, הוכן מאגר נתונים גדול שאליו הוכנסו נתונים מעולמות שונים, כולל היסטוריית שלדת הרכב במשך 15 שנים וסקרי שירות ומכירה שמתבצעים באופן שוטף בחברות הללו. עוד הוכנסו למאגר משתנים של סדרות עיתיות הקשורות למשתני מזג אוויר, ונתונים מבוססי מקום הקשורים למיפוי גיאוגרפי ומיפוי תחרות. מאגר הנתונים היה מספיק עשיר על מנת לתמוך גם בהחלטות חיזוי טקטיות של איתור קהלי יעד (Targeting), כגון זימון לנסיעת מבחן למודל חדש, מניעת נטישת שירותים במוסכים המורשים והגדלת ציי רכב, וגם לתמוך בהחלטות אסטרטגיות. להלן נדון בשני יישומים מתעשיית הרכב שבהם נעשה שימוש במודלים של חיזוי על מנת לתת מענה למספר רב של סוגיות עסקיות ושיווקיות.

ביישום הראשון החברה ביקשה לפתח מודלים לאיתור קהלי יעד עבור שני סוגים של רכבים מסחריים המכונים LCV (Light Commercial Vehicle). אלה כלי רכב מסחריים נפוצים, אחד דגם קטן יותר והשני דגם סטנדרטי. לשם פשטות נכנה אותם רכב מסחרי קטן ורכב מסחרי גדול. מכיוון שהחברה רצתה לבסס גם מסע פרסום (ATL) על סמך המודלים, הוחלט להריץ שלושה מודלים של חיזוי: מודל לכלל הלקוחות שרכשו רכב מסחרי בעשר השנים האחרונות; מודל ללקוחות שרכשו את הדגם הגדול יותר; ומודל לאלו שרכשו את הדגם הקטן יותר. על אף שהמודלים נבדלו זה מזה, כלל הרכבים, הן של הלקוחות הפרטיים והן של הלקוחות המסחריים, הצביעו על אופי שימוש מסחרי. גם משתנים כמו קילומטרז', סוגי ביטוח, סוגי הטיפול במוסכים וימי הטיפול במוסך היו די דומים בכל הדגמים. אם אנשי השיווק היו מסתמכים רק על תוצאות החיזוי, יש להניח שכלל הלקוחות היו מקבלים את אותה הצעה.

עם זאת, נמצאו לא מעט משתנים שרק באמצעות ניתוח לעומק ניתן היה לזהות עד כמה הם משמעותיים. למשל,

מצאנו כי לדגם הקטן יש צבר של מאפיינים (Features) ייחודיים, לדוגמה אינטראקציה מובהקת ובמיקום גבוה בין סוג היישוב (עיר, פרבר וכפר) למגדר. כמו כן, משתנה האינטראקציה בין יישוב עירוני ו"נקבה" נכנס למודל של הרכב הקטן עם מקדם מובהק מאוד. כאשר ביקשנו לבדוק מה מאפיין את הנשים העירוניות בהשוואה לגברים עירוניים שרכשו דגם זה, קיבלנו פרופיל שונה בתכלית. מתברר שהנשים רכשו אביזרים הקשורים לבטיחות ילדים יותר מאשר גברים. מכאן זיהינו מגמה שנשים מעורבות יותר ויותר בבחירת רכב מסחרי שמשרת גם את המשפחה. למותר לציין שרכבים אלה אינם ג'יפים או SUV – אלא רכבים שתכליתם פעילות עסקית זעירה. לאור ממצאים אלה, הוחלט בחברה למצב את המכונית כ"ג'יפ קטן". הופקו שלושה סוגי סרטונים הקשורים למעגל חיי משפחה. בראשון, בחור מחזר אחרי בחורה ומגיע לאסוף אותה בג'יפון. הסרט השני התמקד ביכולת המגוונת של המכונית לשמש ביום כרכב לשילוח סחורה ובערב לבילוי משפחתי, המשרת בעיקר סגמנט זוגות ללא ילדים (DINKIES-Double Income no Kids). בסרט השלישי המשפחה גדלה, והמכונית הוצגה כמכונית בטוחה (Safe car) לנסיעה עם תינוק.

נוסף על פרסום מסיבי בטלוויזיה, ברדיו, בשלטי חוצות ובעיתונים, נשלחו הזמנות למבחני נהיגה על בסיס המודל. מכירת הסגמנט הזה חצתה את כל התחזיות. חברת האם החלה לייצר יותר דגמים "רומנטיים" למכונית, וגרפה פרסים בכל תחומי הפרסום והשיווק. כמובן שתוצאה זו לא הייתה יכולה להתממש בלי ההחלטה של הנהלת החברה על החדרת נתוני עתק, שאפשרה לבצע ניתוח לעומק של המאפיינים של הנשים העירוניות.

ביישום השני, חברת הרכב ביקשה לדעת מהי תפיסת המותג של הרכב ואם אפשר להסיק מסקנות לגבי נאמנות למותג ונאמנות לסוכנות (Dealer). נושא תפיסת מותג נחקר בדרך כלל על ידי סקרים המייצגים מדגם מקרי, קטן יחסית, של האוכלוסייה. משתתפי הסקר נדרשים לענות על מספר שאלות רלוונטיות לבעיה הנחקרת. בדרך כלל השאלות בסקר הם משתנים "איכותיים", שמשקפים תחושות ורגשות (Attitudinal), כגון סקרי שביעות רצון שבהם משתתפי הסקר נדרשים לענות על שאלות של שביעות רצון ממוצר או שירות מסוים (על פי רוב בסולם אורדינלי). להלן נתייחס לנתוני הסקר כאל נתונים "רכים", להבדיל מהנתונים בבסיסי הנתונים של הארגון שנכנה בשם נתונים "קשים". הבעיה עם סקרים היא שהם מתייחסים רק למשתתפי הסקר, המהווים מדגם קטן מהאוכלוסייה. השאלה המרכזית היא איך מרחיבים את תוצאות הסקרים לכלל האוכלוסייה, ובמקרה שלנו: איך אנחנו אומדים את תפיסת המותג (שהוא משתנה "רך") עבור כלל הלקוחות בבסיס הנתונים, על סמך נתונים של סקר שמתבסס על קבוצה קטנה מאוד באוכלוסייה? הרעיון המרכזי הוא למצוא מהם המשתנים ה"קשים" בבסיס הנתונים המתואמים באופן מובהק עם תפיסת המותג על פי הסקר, ואז להשתמש במשתנים האלה על מנת לאמוד את תפיסת המותג בקרב כלל האוכלוסייה. לשם כך יצרנו מדגם אימון שכולל את כל משתתפי הסקר, הוספנו לכל לקוח את מספר השלדה של הרכב שברשותו ומשתנים קשים רלוונטיים נוספים, וכן את תשובותיו לשאלות בסקרים (סקרי קנייה וסקרי שירות). בנינו מודלים של רגרסיות על מנת למצוא אילו משתנים בבסיס הנתונים מתואמים באופן מובהק עם תפיסת המותג. המשתנה התלוי במודלים אלה הוא תפיסת המותג מתוך הסקר שאנחנו מנסים "להסביר" באמצעות המשתנים הרכים והקשים בקובץ האימון. ממצאי המודל הצביעו על כך שתפיסת המותג מהסקר נמצאה עם מתאם גבוה בקרב לקוחות שרכשו יותר מרכב אחד ובקרב לקוחות שנוטים להשתמש במוסכים המורשים. לאור זאת, בנינו ציון מחושב שמשקף את עוצמת המותג המסתמך על הממצאים האלה. למשל, מי שקנה בעבר רכב מאותה חברה וחזר לקנות שוב, או נענה לנסיעת מבחן, נתפס כבעל תפיסת מותג גבוהה. את הנוסחה הזו הפעלנו על כל הלקוחות בבסיס הנתונים על מנת לאמוד את תפיסת המותג עבור כל לקוח בבסיס הנתונים.

כשאנו "מצוידים" באומדים של תפיסת המותג בקרב כלל האוכלוסייה, נסללה הדרך לבדוק מהם המשתנים האחרים בבסיס הנתונים המתואמים באופן מובהק עם תפיסת המותג. אחד הממצאים המעניינים הוא שתפיסת המותג נמצאה במתאם גבוה עם רכישת ביטוח מורחב. למסקנה זו משמעות שיווקית חשובה, שכן משתמע ממנה שכדאי לחברת הרכב להרחיב את מכירות הביטוח המורחב לכלל הלקוחות באוכלוסייה בעלי תפיסת מותג גבוהה שעדיין לא רכשו ביטוח כזה. נציין שמכירת ביטוח מורחב נחשבת לאחד מה-upsell המשמעותיים ביותר בקרב סוכני הנסיעות ויצרני הרכב.

בנוסף, התגלה קשר שלא צפינו מראש בין תפיסת מותג חיובית ובדיקת חורף מוקדמת. כמו כן, נמצאה גם אינטראקציה מובהקת מסדר שלישי, שלפיה מי שקנה אחריות מורחבת וגם התמיד להגיע לביקורת חורף – החזיק בתפיסת מותג גבוהה. לאור הממצאים הללו החליטה חברת הרכב להציע ללקוחות אלה שירות תחת המוטו "אנו דואגים לבטיחותך", והזמינה אותם לקבל שירותים (לאו דווקא בתשלום) הקשורים בתחושת הביטחון ברכב. מבצעי השיווק הותנעו בחורף (לקראת השלגים) ובקיץ (לקראת החופשות והיציאה לבתי קיט). ניתוח הממצאים (Post-mortem) הצביע על אחוז היענות גבוה. השם שניתן לסגמנט הזה היה "האחראיים", ובמערכת ניהול הלקוחות ניתנה להם עדיפות בשירות ובכל השיחות איתם הודגשה רמת הביטחון של הרכב (Safety) והיכולת להיענות מהר לצרכים שונים (Guarantee). בין היתר, הרצנו מודל למציאת השכנים הקרובים ביותר (Nearest Neighbours) לקבוצת "האחראיים" על מנת לאתר לקוחות דומים ולהציע להם לרכוש אחריות מוגדלת.

יותר מזה, ניתוח לעומק של הלקוחות "האחראיים" הצביע על כך שמרביתם קנו את הרכב מיד שנייה, כלומר לא ישירות מהחברה. להנהלת החברה זה היה ממצא חדש, שהצביע על כך שלקוחות שנאמנים למוסכים המורשים לאו דווקא נאמנים לדילרים. ממצא זה חִייב את החברה למצוא דרך שבה הנאמנות למוסך תתורגם לרכישת כלי רכב חדשים. ואכן, הנהלת החברה ביצעה ניתוחי עומק, הגדירה קבוצות מיקוד על בסיס המודל ולמדה את המאפיינים שלהן, העריכה את שווי הלקוחות ובנתה להם תוכנית שימור ושיווק ייחודית.

אחד המשתנים המשמעותיים ביותר בעולם הרכב הוא ערוץ המגע עם הלקוחות. במקרה הזה נמצא כי קבוצת "האחראיים" פונה מספר רב של פעמים למוסכים ולמרכזי השירות וכמעט שלא נוטה להשתמש בערוצים ישירים. המסקנה המתבקשת הייתה שיש להפוך את המוסכים המורשים לערוץ מכירה עבור לקוחות אלו, מה שחייב את החברה לבנות תוכנית שיווק עבור הלקוחות הללו, שכללה הכשרה של נותני השירות להציע ללקוחות דגמים חדשים.

נציין שה"מבט מעבר למודל" ביישום של תפיסת המותג הביא ל"שרשרת" של תובנות בעלי משמעות עסקית. המודל הבסיסי התבסס על נתוני הסקר. ממנו גזרנו את הנוסחה שאפשרה לאמוד את תפיסת המותג לכלל הלקוחות באוכלוסייה, ובכך נסללה הדרך לבנות מודלים נוספים שבהם תפיסת המותג היה המשתנה התלוי, והם הצביעו על תהליכים שיווקיים מעניינים.

בנק השקעות

האירוע השני מתמקד בבנק השקעות גדול במדינה באירופה (שבשל מגבלת סודיות נכנה אותו "הבנק"), שהוא למעשה זרוע של בנק מוביל במדינה. סף הכניסה לניהול תיקים עמד על מאה אלף דולר, והמטרה הייתה לבחון מודלים שונים של לקוחות המנבאים מוכנות לנטילת סיכונים. כחלק מניהול הסיכונים, בנק ההשקעות הזה התמיד להמליץ ללקוחותיו להשקיע במספר מניות המכונות "מניות ברזל". אלו מניות מתחום הבנקאות, האנרגיה, התקשורת, ענף הבנייה, תעשיית זכוכית וכימיקלים – כולן של חברות מובילות באותה מדינה. אחת מהמניות הללו היא של בנק "האם" שנחשב לבנק היציב ביותר במדינה.

גם במקרה הזה נדרש לבנות מאגר נתונים על מנת שניתן יהיה לבנות מודלים של חיזוי. המאגר כלל את תיקי הלקוחות, וכן מידע מפורט על מחיר, תנודתיות והיקף המסחר במניות ובאג"ח. בנוסף שולבו נתונים של קרנות, סלי-מטבעות, שערי מטבעות וכל מוצר בנקאי שהבנק מכר או ייעץ לגביו. השאלות העסקיות התמקדו בתחומי גיוס לקוחות חדשים ובשימור ערך לקוח. אחד הפלחים המעניינים הוא פלח הלקוחות שהחזיקו מספר מניות המוכרות כ"מניות ברזל" בתיקי השקעות, שאחת המרכזיות שבהן הייתה מניית בנק ההשקעות ובנק האם שלו. מודל חיזוי שבנינו עבור סגמנט זה הצביע כי למניית "הבנק" יש תרומה מרכזית בחיזוי רמת הסיכון בתיק ההשקעה, וככל שבתיק ההשקעות של הלקוח היו יותר מניות של בנק האם – כך רמת הסיכון של תיק הלקוח הייתה נמוכה יותר. עם זאת, יש לזכור כי מניה זו הייתה רק אחת מ"מניות ברזל" שבהן משקיעים נהגו להשקיע כדי לגדר את הסיכון.

מפלח זה בודדנו את תת הפלח של לקוחות שהחזיקו את מניות "הבנק" בשיעורים נמוכים, וקראנו לו פלח של "אי החזקת מניות הבנק". הפלח הזה, שהיה בשיעור של אחוזים בודדים בקרב הלקוחות, סקרן את ההנהלה במיוחד מפני שהוא הצביע על איתות של נטישה. על מנת להבין את

מניעי ההשקעה של פלח הלקוחות הזה, ביצענו בדיקה לעומק של מאפייני הלקוחות, כולל בדיקה היסטורית של תיקי ההשקעות שלהם. התברר שלקוחות אלה השקיעו בעבר את מיטב כספם במניות הבנקים, והפסידו הון עתק במשבר של 2009-2008. הם ראו את "הבנק" כאשם לכאורה במצב ולכן צמצמו את ההחזקות שלהם במניות "הבנק".

תהליך זה הציף קבוצה שנפגעה ואיבדה אמון ב"בנק". על מנת למנוע נטישה של הלקוחות האלה יצרה ההנהלה קשר מיידי עם עשרות אלפי הלקוחות הללו והציעה להם להשתתף בהנפקה ייחודית בהנחה גדולה. למותר לציין שבסופו של דבר הלקוחות הללו לא נטשו את "הבנק", ובבדיקה בדיעבד (פוסט-מורטם) התברר שהם אפילו העלו את רמת ההשקעות שלהם.

שוב נציין שאירועי quick win מסוג זה מתאפשרים בעיקר כאשר ניתן לבחון לעומק את כל המשתנים שנכנסו למודלים, תוך שימוש בפלטפורמות ובתוכנות פתוחות שמציגות את משוואות החיזוי בצורה שמאפשרת לבחון את מידת ההשפעה והמדרוג של המשתנים שנכנסו למודל. למעשה מדובר כאן על ניתוח "שאריות" שמתמקד בהקטנת טעות החיזוי באמצעות זיהוי של קבוצות לקוחות ייחודיות ובניית אסטרטגיה שיווקית מתאימה עבורם. במקרה הזה "שלפנו" מתוך פלח הלקוחות את אלה שהחזיקו באחוז נמוך של מניות "הבנק" (מה שהצביע על סבירות של נטישה), והצענו להם תוכנית שימור מתאימה.

תקשורת סלולרית

אחת הבעיות המאתגרות ביותר בחברות טלקום היא לשמר ולהגדיל את רווחיות הלקוחות המשלמים מראש (Prepaid). לקוחות אלו טוענים את מכשיר הטלפון אחת לתקופה, בהתאם לצרכים שלהם, בניגוד ללקוחות המשלמים באמצעות חשבוניות (Postpaid). הלקוחות המשלמים מראש הם בדרך כלל אנונימיים מכיוון שאין עליהם מידע דמוגרפי, ולא תמיד המשלם הוא זה שמשתמש במכשיר. למשל, הורה שמשלם לילד או מעביד שמשלם לעובד.

כסגמנט, לקוחות אלו מהווים תמיד אניגמה למנהלי השיווק שמחויבים להשקיע מאמצים רבים על מנת לשמר אותם ולהעלות את השווי שלהם, במיוחד מאחר שלקוחות אלו לא נוטים להיות "נאמנים" לחברת הסלולר (למותג) אלא נוטים לחפש עסקאות אופטימליות לפני כל טעינה. ואומנם מחלקות השיווק מריצות עשרות מודלים של חיזוי ומודלים למניעת ירידת ערך הלקוח. מכיוון שלקוחות המשלמים מראש אינם מחויבים בחוזה עם חברות הסלולר, לא מקובל לבנות עבורם מודלים למניעת נטישה אלא להציע להם תוכניות שימור כדי לעודד אותם להמשיך ולטעון את המכשיר בחברה, כגון תוכניות תמחור (Price plans), הצעות משולבות (Bundles) ותוספות (Add-ons). הבעיה העיקרית עם לקוחות המשלמים מראש היא שאין לנו עליהם מידע פרסונלי שיאפשר למנהלי השיווק להתאים את תוכניות התמחור ללקוח ואת התוספות שיש להציע לו. כך למשל, אין טעם להציע ללקוחות שלא משתמשים בגלישה סלולרית (Mobile Internet) לרכוש אינטרנט סלולרי (Data). עבור לקוחות דיגיטליים (לקוחות שיש להם טלפון חכם), לא ברור מה תמהיל ההצעה הרצויה בין זמן אוויר (דקות שיחה יוצאות) לבין גלישה סלולרית. מציאות זו מחייבת לחשוב "מחוץ לקופסה" כדי להסיק מהם המשתנים ההתנהגותיים של הלקוחות המשלמים מראש, על סמך ניתוח המשתנים המסבירים שנכנסו למודל החיזוי.

להלן נציג תוצאות של מודלים לחיזוי עבור חברת טלקום אירופית מובילה שמחצית מלקוחותיה משלמים מראש. מטרת המודלים הייתה לענות על כלל השאלות העסקיות העוסקות בהארכת משך חיי הלקוחות (שימור), העלאת הרווחיות, וניהול האסטרטגיה של מעגל חיי לקוח. עולם התקשורת הסלולרית מאופיין במספר גדול של נתונים מ"עולמות תוכן" שונים, ובהם טעינות, שווי לקוחות, תוכניות מחיר, שימוש בחבילות, ניצול זמן אוויר, זמני שימוש (Time bands), ועוד.

את עולמות התוכן האלה תרגמנו למשתנים מסבירים, ובנינו מודלים של חיזוי עבור שני סוגים של בעיות עסקיות:

1. העלאת סכום הטעינה הבודדת (העלאת הכנסות).
2. הסטת לקוחות לא דיגיטליים לדיגיטליים (ניהול מעגל חיי לקוח).

אחד הממצאים הבולטים שהתקבלו מהמודלים האלה הוא שהפעילות הסלולרית בשעות 20:00-18:00 היא אינדיקציה מהותית להבנת הלקוחות, מכיוון שמשתני

פעילות רבים (דקות אוויר יוצאות, דקות אוויר נכנסות, מספר הודעות SMS, ועוד) היו קשורים באופן ישיר או באמצעות אינטראקציה להתנהגות הלקוחות בשעות אלו.

עובדה זו הובילה אותנו לבצע ניתוח לעומק של משתנה שעת היממה (Time band) ברמת הלקוח הבודד. חילקנו את היממה ל-12 "רצועות זמן" (המקבילות לסדר היום) וסיווגנו את רמת הפעילות של כל לקוח ברשת הסלולרית בקבוצת שעות זו בהשוואה לשעות פעילות אחרות. בנוסף, יצרנו משתנים מחושבים הקשורים בחלוקת היממה: בוקר, לפני הצהריים, צהריים, אחר הצהריים, ערב, לילה, ושעות לילה מאוחרות.

המידע הזה אפשר לנו להבחין בין צעירים למבוגרים, בין מי שנמצא בשוק העבודה ומי שלא נמצא בו, לחזות רקע סוציו-אקונומי וכדומה. לדוגמה, ניתן היה לאפיין תלמידים לפי שימוש מועט בטלפון הנייד בשעות הבוקר, ועלייה משמעותית בשימוש בשעות אחר הצהריים והלילה ובסופי שבוע. ללקוחות אלו הייתה גם רשת חברתית גדולה והם נטו להיות יותר דיגיטליים. מבחינת טעינות, ניכר היה שהם מוגבלים בסכומי הכסף ברמה חודשית, והשונות בין הטעינות הייתה קטנה, אם כי סכום הטעינה היה גבוה יותר מהממוצע. לעומתם, לקוחות מבוגרים נוטים להשתמש בטלפון הנייד בשעות הבוקר המוקדמות, מרבים להתקשר לקווים נייחים (מכשירי טלפון בבתים ובמוסדות), מעגל החברים שלהם מצומצם ואין שימוש בשעות לילה מאוחרות. הם גם טוענים את הטלפון בממוצע פעם בשישה שבועות ובסכום קבוע.

ניתוח לעומק של הממצאים בשעות 18:00-20:00 הצביע על מספר קבוצות מובהקות שמאופיינות בהתנהגויות שונות. לדוגמה:

1. בדרך הביתה – קבוצה שאופיינה במספר רב של תאים סלולריים אחרי השעה 18:00 לעומת שעות מוקדמות יותר או מאוחרות יותר. מספר התאים הסלולריים היה גבוה בצורה משמעותית והצביע על העובדה שהלקוחות נמצאים בתנועה ברכבם או בתחבורה ציבורית.

2. לקוחות שאינם מדברים או גולשים בשעות אלו וממעטים בשימוש באופן כללי במכשיר הסלולרי. הנחנו שאלו לקוחות שמאופיינים בטיפול בילדים או עוסקים בפעילות ספורטיבית לאחר שעות העבודה.

3. משכימי קום – לקוחות שממעטים בשיחות בערב אך מתקשרים וגולשים באופן משמעותי בשעות הבוקר המוקדמות מאוד. זוהי תכונה שמאפיינת בדרך כלל אנשים מבוגרים.

4. תלמידים/צעירים – מאופיינים בשימוש גבוה בשעות 18:00-20:00, במיעוט שיחות וגלישה סלולרית בשעות הלימודים, ובשימוש משמעותי בסופי שבוע.

דווקא משום שהלקוחות אנונימיים, ניתוח המשתנה הזה, שנמצא משמעותי בכל המודלים, הוביל להחלטות לגבי זמני פנייה ללקוחות, וכן שפה שיווקית ואסטרטגיה לניהול חיי הלקוח. לדוגמה, תלמידים סומנו לקראת מבצעי חופש גדול וחזרה לבית הספר. האימהות סומנו כמועמדות לתוכניות (postpaid) כדי "להגן על המשפחה".

ומה קורה במודלים סגורים?

כאמור, אי אפשר לסיים מאמר הדן באינטרפרטציה של מודלים לחיזוי בלי להתייחס גם למודלים סגורים, במיוחד כשחלק גדול מהמודלים לחיזוי בעולם נתוני העתק, אם לא רובם, הם מודלים סגורים. גם במודלים סגורים מקובל להעריך את טיב החיזוי באמצעות תיקוף המודל על קובץ תיקוף שלא השתתף בתהליך בניית המודל, ולהשוות בין הערכים האמיתיים של המשתנה התלוי מול הערכים החזויים ממודל החיזוי.

אלא ששיטת תיקוף זו היא שיטה גלובלית להערכת טיב החיזוי של מודל החיזוי. מתבקשת שיטה לניתוח פרטני הכוללת מדדים שמאפשרים להעריך את ההשפעה של המשתנים המסבירים על משתנה התוצאה (המשתנה התלוי) גם ברמת התצפית הבודדת, בדומה לדיון הפרטני לעיל עבור מודלים פתוחים. מדדים כאלה נדרשים לעיתים על ידי גופים רגולטוריים או אפילו על ידי הלקוחות עצמם כדי להבין את תוצאות המודל, במיוחד כשמדובר על תחזיות שגויות שכרוכות בנזק רב או בעלות גבוהה. למשל, להסביר מדוע בקשתו של אדם מסוים לקבל הלוואה אושרה אך בקשתו של אדם אחר נדחתה, להסביר מדוע לקוח מואשם בהונאה, ועוד. נציין שבמודלים פתוחים כמו רגרסיה ליניארית, ההשפעה של המשתנה המסביר על משתנה התוצאה מתקבלת באמצעות המכפלה של המקדם של

המשתנה במודל הרגרסיה בערך של המשתנה המסביר. בעיה זו מסתבכת כשמדובר במודלים סגורים. התחום של בינה מלאכותית ברת-הסבר (explainable AI) מציע מספר מדדים לכמת את יכולת ההסבר של משתנים גם במודלים סגורים של חיזוי, שהנפוצים שבהם הם מדד LIME וערך שפלי (Shapley value). שני המדדים האלה הם Model-agnostic, כלומר ניתן ליישם אותם על כל מודל של חיזוי.

מדד LIME (Local Interpretable Model-agnostic Explanation)

גישת LIME (Ribeiro et al., 2016) מציעה שיטת תיקוף מקורבת של התחזיות עבור מודלים "סגורים". הרעיון הוא ליצור מדגם "לוקאלי" עבור תצפיות ספציפיות "בעלות עניין" (למשל תצפיות חריגות) שכוללות מספר תצפיות שכנות, או אפילו "לשתול" מספר תצפיות מלאכותיות בעלות מאפיינים דומים מסביב לתצפית הספציפית, ואז לבנות מודל חיזוי פשוט (למשל מודל ליניארי) עבור המדגם הלוקאלי ולהשתמש במקדמים של המודל הליניארי כדי לאמוד את התרומה "המקומית" של המשתנים המסבירים למודל. הנחה העבודה של גישת LIME היא שהתרומות של המשתנים המסבירים של גישת LIME מהווים קירוב מספיק טוב של המודל הסגור עבור תצפיות עם מאפיינים דומים.

ערך שפלי (Shapley Value)

בניגוד לגישת LIME שנותנת מדד מקורב לתרומה של המשתנים השונים לתחזית של המודל, ערך שפלי הוא מדד מדויק להערכת התרומה של משתנה מסביר לאיכות החיזוי, ויש לו גם בסיס תיאורטי מוצק המתבסס על תורת המשחקים השיתופיים (Shapley, 1953). הרעיון הוא שכל המשתנים המסבירים במודל חיזוי משתפים פעולה על מנת לחזות את משתנה התוצאה. אך התרומה השולית של כל משתנה מסביר לחיזוי התוצאה שונה לא רק ממשתנה למשתנה, אלא גם מהרכב המשתנים המסבירים שנכנסים איתו למודל. קיימות מספר קומבינציות שונות להכניס משתנה מסוים למודל. כל קומבינציה כזו היא "קואליציה". ערך שפלי עבור משתנה מסביר מוגדר בתור הממוצע של התרומות השוליות של המשתנה עבור כל הקואליציות האפשריות. כדי לחשב את התרומה השולית של משתנה עבור קואליציה כלשהי, מפעילים את מודל החיזוי על הקואליציה כדי לאמוד את הערך של משתנה התוצאה עם ובלי המשתנה, ומחשבים את ההפרש בין התחזיות.

בתוקף הגדרתו בתור ממוצע התרומות השוליות של משתנה מסביר, ערך שפלי מבטא את החשיבות של המשתנה בתהליך החיזוי. ככל שערך שפלי של משתנה מסביר גבוה יותר ובסימן חיובי, הוא תורם יותר להסבר של משתנה התוצאה עבור תצפית נתונה (וההיפך לגבי ערך שפלי שלילי). לכן ניתן להסביר באמצעותו את תוצאות החיזוי אפילו ברמה של התצפית הבודדת. בלי להיכנס לפרטים, נציין שלערך שפלי מספר תכונות שמבטיחות חלוקה "הוגנת" של התרומות של המשתנים המסבירים, מה שהופך אותו לקריטריון המועדף להערכת התרומה של המשתנים המסבירים במודלים של חיזוי.

החיסרון של ערכי שפלי הוא כמות החישובים הרבה שעולה באופן אקספוננציאלי כפונקציה של מספר המשתנים המסבירים. עם זאת, קיימים מספר קירובים של ערכי שפלי, שנדונים בספרות, ומאפשרים חיסכון רב בחישובים.

סיכום

מאמר זה עוסק ביתרונות ובאופן שבו ניתן להגיע לתובנות עסקיות ממודלים פתוחים של חיזוי באמצעות למידה לעומק של מקדמי המודלים ושל האינטראקציות בין המשתנים. בנוסף, אנו דנים בקצרה במדדים שמציע התחום של Explainable AI כדי להעריך את התרומות של המשתנים המסבירים גם במודלים סגורים.

להבדיל ממודלים סגורים שהם שקופים למשתמש, מודלים פתוחים של חיזוי הם מודלים שבהם כל מרכיבי המודל חשופים וידועים למשתמש. בעידן נתוני העתק, הידע החבוי במודלים הפתוחים יכול להצביע על תובנות עסקיות שטומנות בחובן סיכוי להעלאת ההכנסות ושיפור הרווחיות. במאמר זה התמקדנו בעיקר ביכולת להסיק מסקנות ממודלים פתוחים של חיזוי אנליטי, העוסקים בחיזוי שיעורי התגובה של אירועים עתידיים על סמך תצפיות מהעבר שעבורם ערכי התגובה ידועים. תהליך ההפקה של תובנות

עסקיות נוספות ממודל החיזוי מחייב את מדעני הנתונים להביט מעבר למודל, לעיתים תוך שילוב של נתונים נוספים שלא בהכרח נדרשים לצורך בניית מודל החיזוי, על מנת למצות את מלוא התועלת מהמודל. במאמר זה הדגמנו באמצעות שלושה אירועים אמיתיים מתחום שיווק הרכב, בנקאות השקעות וגלישה סלולרית, כיצד ניתן לנצל תובנות הנובעות ממודלים פתוחים של חיזוי אנליטי כדי לבנות אסטרטגיה עסקית ומהלכים שיווקיים.

כאמור, תהליך זה של חשיבה "מחוץ לקופסה" מחייב לעיתים שימוש בנתונים נוספים מעבר לאלו שנדרשים עבור מודל החיזוי. למשל, באירוע הרכב לא להציע רכבים חשמליים ללקוחות שגרים באזורים ללא תשתית הטענה. תובנה זו מתאפשרת רק על סמך הוספה של משתנה המיקום הגיאוגרפי למאגר הנתונים, משתנה שלא בהכרח נדרש עבור מודל החיזוי האנליטי, אבל נדרש בתור "מסנן" כדי להבחין בין אזורים שבהם יש תשתית טעינה לבין אזורים ללא תשתית כזו. לכן תנאי הכרחי לחשיבה מחוץ למודל הוא בניית מאגרי נתונים שמשלבים נתונים ממקורות שונים, לאו דווקא כאלה שנדרשים עבור מודל החיזוי, כולל נתונים מחוץ לארגון. משימה זו דורשת הכרה עמוקה של תחום הבעיה, של מודל החיזוי ושל בסיסי הנתונים הרלוונטיים לארגון (פנימיים וחיצוניים), ומחייבת מעורבות פעילה של מדעני נתונים בכל התהליך.

נציין שתהליך זה של מבט מעבר למודל, המתבסס על מודלים "פתוחים", לא שולל את הצורך במודלים סגורים, שכן מדובר בקהל יעד שונה. מודלים פתוחים מיועדים עבור מדעני הנתונים ויועצים למיניהם, שיש להם את היכולות, הידע והכלים להפיק מסקנות עסקיות חשובות על סמך ניתוח לעומק של מאפייני המודל. מודלים סגורים מיועדים עבור המשתמשים העסקיים (business users), שאין להם את היכולות, המשאבים והמיומנויות לבנות מודלים רב-ממדיים לחיזוי אנליטי, אבל הם עדיין זקוקים למודלים האלה לצורך קבלת החלטות וקידום הפעילות העסקית שלהם. לכן אין כאן שאלה של אילו מודלים לפתח – סגורים או פתוחים – מפני שלכל אחד מהם יש תפקיד שונה. המצב האידיאלי הוא לפתח מודלים "היברידיים" שהם גם סגורים וגם פתוחים בו-בזמן. המשתמש העסקי יכול להשתמש בערכי ברירת המחדל של המודל על מנת לבנות מודלים של חיזוי בצורה אוטומטית ולגשת ישירות לחישוב הציונים של לקוחות חדשים. במקביל מדעני הנתונים ו/או משתמשי קצה סקרנים יוכלו "לפתוח" את המודל על מנת להפיק ממנו תובנות עסקיות נוספות.

פיתוח מודלים "היברידיים" מהסוג הזה מציב אתגרים חדשים ומעניינים עבור החברות העוסקות בפיתוח תוכנות בתחום החיזוי האנליטי, מה שמסביר מדוע מודלים היברידיים מסוג זה עדיין נדירים יחסית. עם זאת, מספר חברות כבר מפתחות מערכות כאלו כיום, והציפייה היא שמגמה זו תצבור תאוצה עם הגידול בביקוש למודלים מהסוג הזה.

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

רשימת מקורות

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

Athey, S. and Guido, I. (2016). Recursive Partitioning for Heterogeneous Casual Effects, *Proceedings of the National Academy of Sciences* 113, 7353-7360.

Friedman, J.H. (1999). *Greedy Function Approximation: A Gradient Boosting Machine*, online PDF document

Gandhi, P. (2019). *Explainable Artificial Intelligence*, https://www.kdnuggets.com/2019/01/explainable-ai.html

Hastie, T., Tibshirani, R. and Friedman, J. (2009). *The Elements of Statistical Learning*, Springer-Verlag, NY.

Pearl, J. (2000). *Causality: Models, Reasoning, and Inference*, Cambridge University Press.

Ribeiro, M.T., Sameer S., and Carlos, G. (2016). Why Should I trust you? Explaining the Predictions of Any Classifier, *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1135-1144.

Rokach, L. (2010). Ensemble-based classifiers, *Artificial Intelligence Review*, 33, 1–39.

Schoenborn, J. M. and Althoff, J. D. Recent Trends in XAI: A Broad Overview on Current Approaches, Methodologies and Interactions, http//gaia.fdi.ucm.es/events/xcbr/papers/XCBR-19_paper_1/pdf

Shapley, S.L. (1953). A Value for n-person Games, in Kuhn, H.W. and Tucker, A.W. (eds). Contributions to the Theory of Games, *Annals of Mathematical Studies*, Princeton University Press, 28, 307-317.

Shmueli, G. (2010). To Explain or to Predict? *Statistical Science*, 25, 289-310.

Shmueli, G. and Koppius, O. (2011). Predictive Analytics in Information Systems Research, *MIS Quarterly*, 35, 553-572.

Wager, S. and Athey, S. (2018). Estimation and Inference of Heterogeneous Treatment Effects Using Random Forests. *Journal of the American Statistical Association*, 113, 1228-1242.