תיקוף ובדיקה

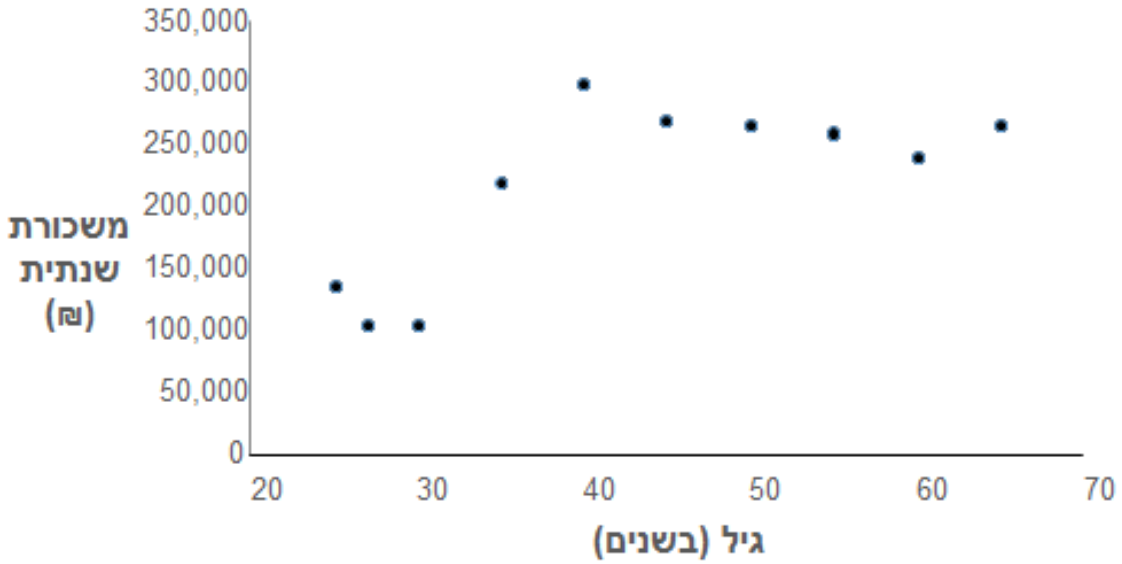
כאשר נתונים משמשים לניבוי או לקביעת אסטרטגיית החלטה, באופן טבעי קיימת סכנה שמודל של למידת המכונה יעבוד טוב על הנתונים ששימשו לפיתוחו, אך לא יכליל (Generalize) בצורה טובה נתונים חדשים. מכאן עולה נקודה ברורה לפיה, חשוב מאוד שהנתונים אשר משמשים לפיתוח מודל של למידת מכונה ייצגו בצורה טובה את המצבים שעליהם ייושם המודל בעתיד. לדוגמא, שימוש בנתונים אודות מכירות, באזור מסוים שבו הלקוחות הינם בעלי הכנסה גבוהה, לצורך ניבוי המכירות עבור מוצר מסוים רק סביר שייתנו תוצאות מוטות.

כפי שסטטיסטיקאים הבינו זה מכבר, חשוב ביותר גם לבדוק את המודל באמצעות מדגם אחר (Out-of-Sample). במילים אחרות, יש לבדוק את המודל על מדגם נתונים שונה מזה ששימש לקביעת הפרמטרים של המודל. מדעני נתונים מכנים את מדגם הנתונים שמשמש לפיתוח המודל בשם "סט אימון" (Training Set), בעוד שאת מדגם הנתונים המשמש לקביעת רמת הדיוק של המודל הם מכנים בשם "סט בדיקה" (Test Set, לעיתים קרובות יש הם גם עושים שימוש בסט תיקוף, Validation Set). במאמר זה נמחיש את השימוש בסט אימון ובסט בדיקה באמצעות דוגמא פשוטה. נניח שאנו רוצים לבנות מודל לניבוי/חיזוי המשכורות השנתיות של אנשים שעובדים באותו מקצוע באזור מסוים רק על סמך הגילאים שלהם. אז לקחנו מדגם נתונים של 10 אנשים (לשם המחשה בלבד לקחנו מדגם מאוד קטן. נעיר כי מערכי הנתונים המשמשים בלמידת מכונה גדולים פי עשרות מונים מהמדגם שלנו). נתוני מדגם זה אשר אותו נכנה "סט האימון", מוצגים בלוח שלהלן ומתוארים בתרשים שלהלן:

להלן נתוני סט האימון: משכורות מתוך מדגם מקרי של 10 אנשים שעובדים במקצוע מסוים באזור מסוים.

גיל (בשנים)	משכורת שנתית (₪)
25	135,000
55	260,000
27	105,000
35	220,000
60	240,000
65	265,000
45	270,000
40	300,000
50	265,000
30	105,000

להלן תרשים פיזור (Scatter Plot) של נתוני סט האימון:

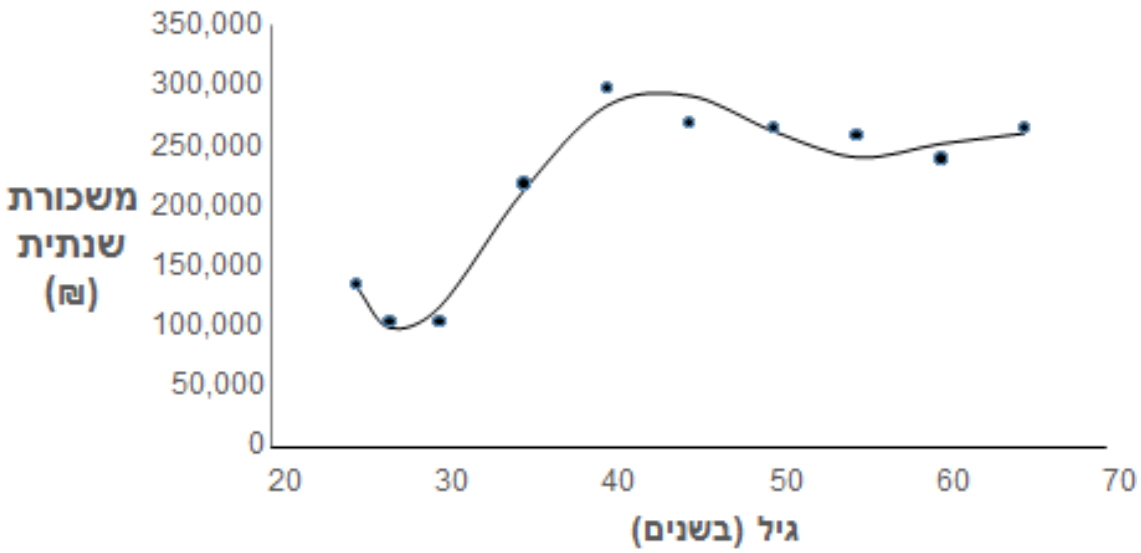


אין דבר מפתה יותר מלפתח מודל שיתאים את עצמו בצורה הטובה ביותר לנתונים הקיימים. מניסיוני כמדען נתונים מודל פולינום מסדר חמישי "עושה את העבודה" כמו שאומרים. למי שלא מכיר, ככה נראה מודל פולינום מסדר חמישי:

$$Y = a + b_1X + b_2X^2 + b_3X^3 + b_4X^4 + b_5X^5$$

כאשר Y הוא המשכורת השנתית ו- X הוא הגיל של מקבל המשכורת. התוצאות עבור רמת ההתאמה של מודל הפולינום מסדר חמישי לנתוני סט האימון מוצגים בתרשים שלהלן.

להלן תוצאות טיב ההתאמה (Goodness of Fit) של מודל הפולינום מסדר חמישי לנתוני סט האימון:



כפי שניתן לראות, מודל הפולינום מסדר חמישי מספק התאמה טובה לנתוני סט האימון. סטיית התקן של ההפרשים שבין המשכורת החזויה על בסיס המודל והמשכורת בפועל עבור עשרת האנשים שסט האימון, מכונה "שורש השגיאה הריבועית הממוצעת" (RMSE- Root Mean Square Error) והיא נאמדה על ידינו בכ- 12,902 ₪.

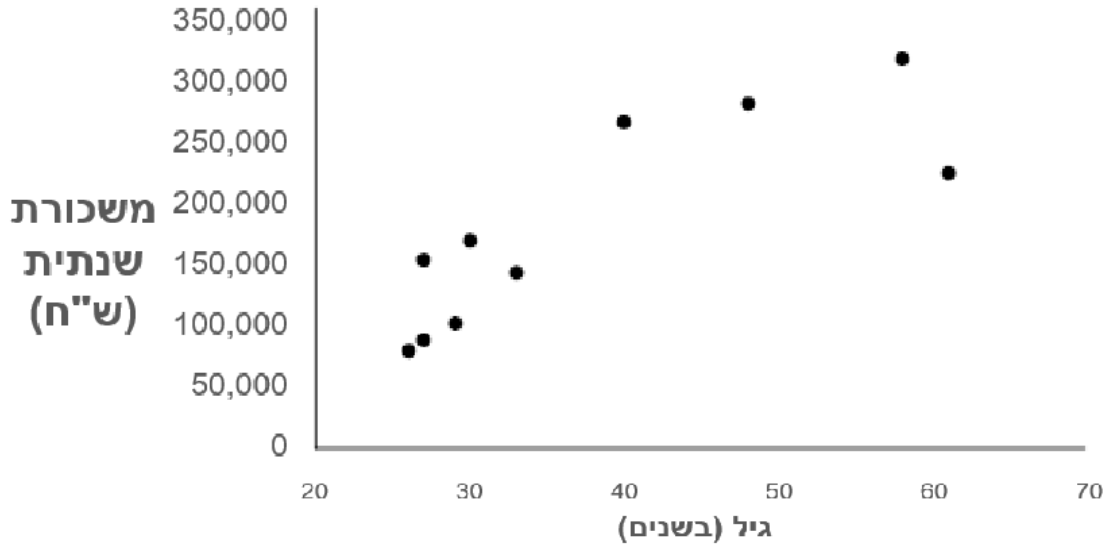
עם זאת, ההיגיון הבריא אומר לנו שיכול להיות שהמודל שלנו מתאים יתר על המידה (Over-Fitts) לנתוני סט האימון וזאת בשל העובדה שהעקום שבתרשים לעיל לא נראה מציאותי בעליל: העקום יורד, עולה, יורד ואז עולה שוב ככל שהגיל עולה. למעשה עלינו לבדוק את המודל באמצעות נתוני מדגם אחר (Out-of-Sample). ואם נעבור משפת האקונומטריקה לשפת מדע הנתונים, אנו צריכים

לקבוע האם המודל מכליל (Generalize) בצורה טובה נתונים חדשים ששונים מהנתונים אשר שימשו לפיתוח המודל.

להלן נתוני סט הבדיקה: משכורות מתוך מדגם מקרי של 10 אנשים נוספים שעובדים במקצוע מסוים באזור מסוים.

גיל (בשנים)	משכורת שנתית (₪)
30	166,000
26	78,000
58	310,000
29	100,000
40	260,000
27	150,000
33	140,000
61	220,000
27	86,000
48	276,000

להלן תרשים פיזור של נתוני סט הבדיקה:



כאשר אנו משתמשים במודל פולינום מסדר חמישי אנו מגלים ששורש השגיאה הריבועית הממוצעת (RMSE) שלו על נתוני סט הבדיקה הוא בערך 38,794 ₪, הרבה יותר גבוה משורש השגיאה הריבועית הממוצעת שלו על נתוני סט האימון (12,902 ₪).

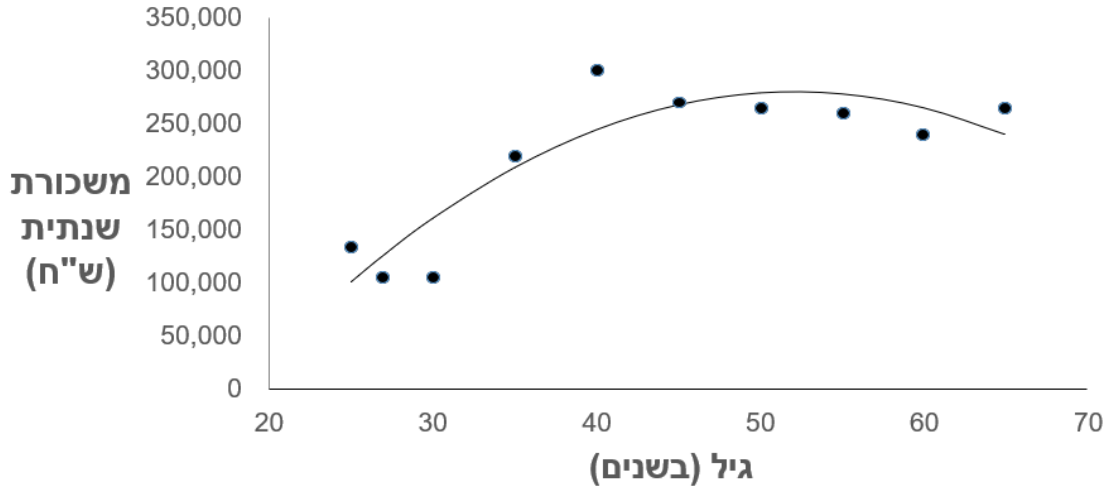
זוהי אינדיקציה ברורה לכך שמודל הפולינום מסדר חמישי מתאים יתר על המידה לנתוני סט האימון, משמע, הוא אינו מכליל בצורה טובה את נתוני סט הבדיקה.

במקרה שכזה, הצעד הטבעי והמתבקש הוא לבחון מודל פשוט יותר, כמו למשל מודל ריבועי:

$$Y = a + b_1 X + b_2 X^2$$

כלומר, פולינום מסדר שני.

להלן תוצאות טיב ההתאמה של המודל הריבועי לנתוני סט האימון:



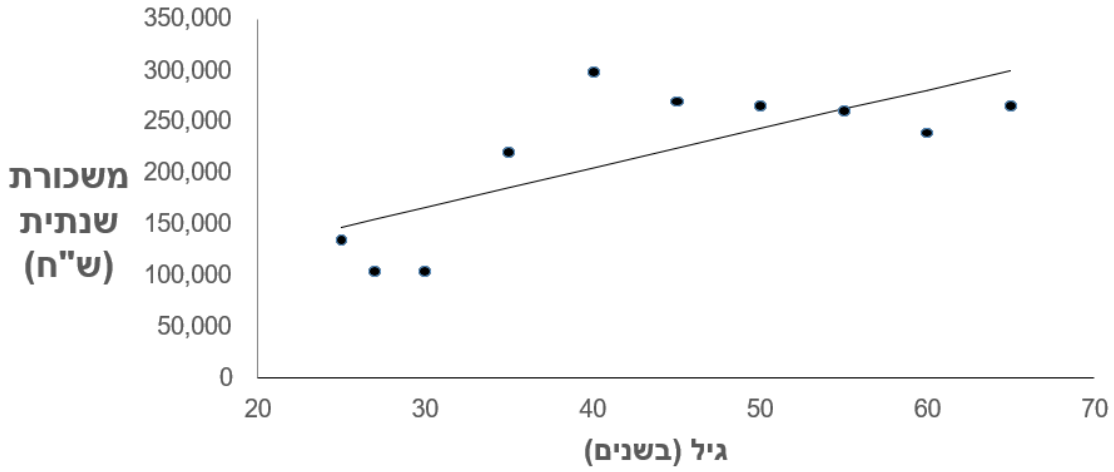
כמובן שההתאמה איננה טובה כמו זו של מודל הפולינום מסדר חמישי. שורש השגיאה הריבועית הממוצעת (RMSE) של המודל הריבועי על נתוני סט האימון הוא 32,932 ₪. עם זאת, המודל מכליל בצורה טובה נתונים חדשים. שורש השגיאה הריבועית הממוצעת (RMSE) של המודל הריבועי על נתוני סט הבדיקה הוא 33,554 ₪, קצת פחות גרוע משורש השגיאה הריבועית הממוצעת שלו על נתוני סט האימון (32,932 ₪). לפיכך, ניתן רק לצפות שהמודל הריבועי יספק תחזיות טובות יותר למשכורת השנתית עבור עובד בגיל מסוים מאשר מודל מורכב יותר כמו מודל הפולינום מסדר חמישי.

כאמור המודל הריבועי הינו פשוט יותר ממודל הפולינום מסדר חמישי וגם עובד טוב יותר ממנו ככלי ניבוי. עם זאת, אין זה אומר בהכרח שמודלים פשוטים יותר לעולם יהיו טובים יותר ממודלים מורכבים. לשם המחשה, הבה נבחן מודל עוד יותר פשוט יותר, כמו למשל מודל לינארי:

$$Y = a + b_1 X$$

כלומר, פולינום מסדר ראשון.

להלן תוצאות טיב ההתאמה של המודל הלינארי לנתוני סט האימון:



מבחינה ויזואלית ניתן לראות כי המודל הלינארי אינו מצליח "לתפוס" את שיעור הדעיכה (או שיעור הצמיחה היורד) של המשכורות עבור אנשים מעל לגיל 50. ממצא זה מקבל אישור באמצעות העובדה ששורש השגיאה הריבועית הממוצעת (RMSE) של המודל הלינארי על נתוני סט האימון הוא 49,731 ₪, הרבה יותר גרוע משורש השגיאה הריבועית הממוצעת שלו על נתוני סט האימון (32,932 ₪).

להלן סיכום תוצאות שורש השגיאה הריבועית הממוצעת של שלושת המודלים עבור שני סטים של הנתונים:

	מודל פולינום מסדר חמישי	מודל ריבועי	מודל לינארי
סט האימון	12,902	32,932	49,731
סט הבדיקה	38,794	33,554	49,990

מהטבלה לעיל עולה כי הן המודל הלינארי והן המודל הריבועי מכלילים בצורה טובה נתונים חדשים, בעוד שמודל הפולינום מסדר חמישי איננו מכליל בצורה טובה נתונים חדשים. לאמור- מודל הפולינום מסדר חמישי מתאים יתר על המידה לנתוני סט האימון, בעוד שהמודל הלינארי מתאים חסר על המידה לנתוני סט האימון.

כיצד אם כך ניתן להשיג את האיזון בין התאמת יתר להתאמת חסר? זוהי סוגיה חשובה בלמידת מכונה. חלק מהאלגוריתמים של למידת מכונה דורשים מספר גדול מאוד של פרמטרים ולכן קל מאוד להגיע למצב של התאמת יתר כאשר סט האימון הוא גדול. בהתבסס על המקרה הפשוט שניתחנו במאמר זה, ניתן לנסח את כלל האצבע הבא:

"יש להמשיך ולהעלות את רמת המורכבות של המודל עד אשר נתוני סט הבדיקה (קרי, נתוני מדגם אחר) מצביעים על כך שהמודל כבר איננו מכליל בצורה טובה נתונים חדשים."

בדוגמת המשכורות שלנו השתמשנו בשני סטים של נתונים. האחד (סט האימון) שימש לפיתוח המודל והאחר (סט הבדיקה) שימש לבדיקת המודל על נתוני מדגם אחר.

באקדמיה, כלכלנים אמפיריים (קרי, אקונומטריקאיים וסטטיסטיקאים) מפתחים תמיד רק מודל אחד ולאחר מכן מחליטים האם הוא שימושי דיו על לביצוע תחזיות, אם לאו. בפרקטיקה, מדעני נתונים (קרי, אקטוארים ומנהלי סיכונים) מפתחים מספר מודלים ומשווים את התוצאות (ממש כפי שאנו עשינו במאמר זה). לפיכך, הפרקטיקה המיטבית הינה לחלק את הנתונים הזמינים לשלושה סטים: 1) סט אימון; 2) סט תיקוף; ו- 3) סט בדיקה.

סט האימון משמש לפיתוח מודלים אלטרנטיביים, בעוד שסט התיקוף נועד פעם אחת לחקור באיזו מידה המודלים שפותחו באמצעות סט האימון מכלילים נתונים חדשים ופעם שניה לבחור מבין המודלים הללו את המודל שמכליל בצורה הטובה ביותר נתונים חדשים. סט הבדיקה "נשמר בצד" כגיבוי ומשמש בתור נתוני מדגם אחר לצורך בדיקת מידת הדיוק של המודל הנבחר בתום התהליך.

חלוקה טיפוסית של הנתונים הינה כדלקמן: 60% לסט האימון, 30% לסט התיקוף ו- 10% לסט הבדיקה. עם זאת, זה תלוי כמובן במודלים ובכמות הנתונים הזמינים.

חשוב ביותר להדגיש שסט האימון שעשינו בו שימוש בדוגמא זו איננו מייצג נורמטיבי לסט אימון בלמידת מכונה (לכל בר בי רב ברור הלא שמעשר תצפיות לא ממש ניתן ללמוד בצורה מהימנה על קשר מסוים). על כן מטרת מאמר זה הייתה להמחיש את בעיית התאמת היתר/התאמת החסר באמצעות דוגמא פשוטה למדי.

<u>פרטים אודות כותב המאמר: מדען הנתונים רועי פולניצר, PDS</u>

- מייסד ומנכ"ל האיגוד הישראלי למדעני נתונים מקצועיים (PDSIA), מייסד ויו"ר לשכת מעריכי השווי והאקטוארים הפיננסיים בישראל (IAVFA) ובעלים של פירמת הייעוץ וההדרכה שווי פנימי.
- מחזיק בתואר M.B.A. במנהל עסקים עם התמחות בניהול סיכונים ואקטואריה ותואר B.A. בכלכלה עם התמחות במימון שניהם בהצטיינות מאוניברסיטת בן-גוריון בנגב, דיפלומה בניהול סיכונים פיננסיים (FRM) מאוניברסיטת אריאל, תואר Financial Risk Manage מארגון בינ"ל GARP, תואר Certified Risk Manage מארגון ישראלי IARM, תואר Fellow Actuary מארגון ישראלי IAVFA ותואר Professional Data Scientist מארגון ישראל PDSIA.
- בעל ניסיון אינטנסיבי של מעל עשור וחצי שנים בתחום מדע הנתונים ולמידת המכונה, הכולל ביצוע מחקרי מידע מעמיקים לשם הפקת תובנות עסקיות, ניקוי, טיוב וסידור של המידע המשמש למחקרים השונים, הפעלת אלגוריתמים שונים של מידול, כריית נתונים ו- Machine Learning על המידע ובניית תהליכי הכנת המידע והאופטימיזציה של האלגוריתמים השונים.
- מרצה לתכנות בשפות R ו- Python, לניהול סיכונים, הערכות שווי ואקטואריה והנדסה פיננסית.