

# רוצה לדעת מתי הלקוחות עומדים לעזוב אותך? תתייעץ עם מדען נתונים

**רועי פולניצר** מסביר מדוע מדען נתונים מבין יותר בסטטיסטיקה ואקונומטריקה מכל איש מדעי המחשב או מהנדס תוכנה וגם טוב יותר בתכנות מכל סטטיסטיקאי או כלכלן



לך חודש לפני? שבועיים לפני? יום לפני? אלו דברים שצריך להגדיר אותם היטב וכמו גם את המדדים שלהם.

## שלב 2

בשלב השני מדען הנתונים אוסף את הנתונים, "מנקה", מכין ומעבד אותם. בעיה אחת שמאפיינת את השלב הזה היא בעיית המידע החסר. בבעיה זו אני נדרש להחליט האם אני משלים את המידע החסר או פשוט זורק את המידע הלא שלם. לדוגמה, נניח שיש לי עמודת נתונים מסוימת שחסרים בה הרבה נתונים, האם אני יכול לוותר על כל העמודה הזאת או שלא?

בעיה אחרת היא בעיית החריגים, כאשר חריגים אלו מקרים מאוד קיצוניים. לדוגמה, נניח שאני מנסה להבין משהו על הלקוחות של סוכן ביטוח מסוים שרוב הלקוחות שלו משלמים פרמיות בסך 300-500 שקל לחודש ופתאום אני נתקל

כלומר, בעיה שאני הולך לפתור באמצעות כלים מתמטיים.

לדוגמה, לפני כשבועיים פנה אליי סוכן ביטוח ושאל אותי על בעיית נטישת לקוחות. אותו סוכן ביטוח רוצה לדעת לזהות מתי לקוחותיו עומדים לנטוש אותו, לפני שזה קורה. כמדען נתונים כשאני מתחיל לתרגם בעיה שכזו למספרים או מתחילות לעלות לי כל מיני שאלות אחרות, כמו למשל: מה זה לקוח? לקוח זה מישהו שנמצא אצלך לפחות שנה? לקוח זה מישהו שמשלם לך כל חודש? זה מאוד תלוי בענף (ביטוח אלמנטרי, ביטוח חיים או ביטוח בריאות).

שאלה אחרת היא מה זה נטישה? האם נטישה זה לקוח שהפסיק לשלם פרמיות? או שנטישה זה רק לקוח שפרדה את הפוליסה? האם נטישה זה אי חידוש? האם נטישה זה לקוח שהודיע לך שהוא החליף סוכן? שאלה נוספת היא מה זה בכלל חיווי נטישה? האם אתה רוצה שאני אבנה מודל שיתריע

דע נתונים הוא תחום שבו מסיקים מסקנות מתוך נתונים. הבעיה שברגע שמסתפקים בהגדרה הצרה הזו או למעשה מעריכי שווי, אקטוארים ומנהלי סיכונים אומרים שהם עושים את אותו הדבר והאמת שהם צודקים. העבודה של הרבה מאוד אנשי מקצוע היום היא למעשה מדע נתונים מכל מיני היבטים. השוני בין מדע נתונים לבין הערכות שווי, ניהול סיכונים ואקטואריה נעוץ במתודולוגיה הסדורה של מדע הנתונים המכונה מתודולוגיית 6 השלבים ל-CISPDPM (תהליך סטנדרטי חוצה ענפים לכריית נתונים).

## שלב 1

על פי מתודולוגיה, השלב הראשון בכל פרויקט מדע נתונים הוא בעצם שלב הבנת הבעיה. בשלב זה המטרה שלי כמדען הנתונים היא לתרגם את הבעיה העסקית הניצבת לפניי לבעיה מתמטית,

צילום ראשי: עמית איליך

## PassportCard Business

הכניסה שלך לעולם העסקי  
בביטוח נסיעות לחו"ל

לפרטים נוספים [pcbusiness@passportcard.co.il](mailto:pcbusiness@passportcard.co.il)



הכרזת

באמצעות פספורטכארד ישראל סוכנות לביטוח כללי (2014) בע"מ, החברה המבטחת - הפניקס חברה לביטוח בע"מ, בכפוף לתנאי הפוליסה, חריגה וסייגיה ותנאי תוכנית PassportCard Business, בכפוף לחיתום רפואי, ללא השתתפות עצמית בשירותים רפואיים בחו"ל באמצעות הכרטיס, וכפוף למגבלות השימוש בכרטיס, נמלות משיכה ומגבלות נקודתיות בכספומטים השונים.

חמשת השלבים הללו, אני מגיע לשלב השישי, הלא הוא שלב שילוב המודל. בשלב זה אני מעביר את המודל ללקוח וממליץ לו היכן עליו למקם את המודל בתהליך קבלת ההחלטות שלו בארגון.

לסיכום, הרבה מאוד אנשים חושבים שמדע נתונים זה רק יישום, פיתוח ובניית מודלים (קרי, החלק של למידת המכונה) ולכן הם נוטים לחשוב שמדע נתונים הוא שם נרדף לסטטיסטיקה. לעניות דעתי מדובר בטעות, הואיל וכיום עולם מדע הנתונים תופס או כל ששת השלבים של מתודולוגיית ה-CISPDM כתהליך שלם ולא רק את השלבים שבהם נכנסת הסטטיסטיקה לפעולה.

הסטטיסטיקה נכנסת רק בשלבים של איסוף הנתונים ובניית המודל. למשל, בשלב איסוף הנתונים אני משתמש בסטטיסטיקה על מנת להגיד שמשתנה מסביר מסוים הוא לא משמעותי או לא אינפורמטיבי מספיק או לחילופין ששני משתנים מסבירים מסוימים אומרים את אותו הדבר ולכן אני יכול לוותר על אחד מהם. בשלב בניית המודל, למשל, אני מכוון את המודלים שאני בונה באמצעות כל מיני הנחות סטטיסטיות או שאני מכוון את המשתנה המוסבר שלי כך שיתפלג בצורה מסוימת, אבל עדיין המרחק בין מדע נתונים לסטטיסטיקה הוא רחוק.

הכותב הינו מדען נתונים ונחשב מומחה בתחומי המימון, ניהול הסיכונים, האופציות וההנדסה הפיננסית

**שלב 5**

שלב אחד לפני שאני מעביר את מודל החיזוי ללקוח, אני נוהג לברוק האם המודל שבניתי ותיקפתי הוא אכן הגיוני. זהו למעשה השלב החמישי במתודולוגיה, שלב בדיקת הגיוניות המודל. מדעני נתונים מקדישים זמן לא מבוטל בלנסות ולהבין מה המודל שהם בנו ותיקפו אומר להם.

לדוגמה, מודל הרגרסיה הליניארית, הנלמד בתואר ראשון בכלכלה, מספר לנו סיפור בצורה מסוימת. מודל עץ החלטה, הנלמד בתואר שני במנהל עסקים, שואל שאלות על מודל הרגרסיה ונותן אינדיקציות אחרות. וכמובן יש את המודלים של "למידה עמוקה", שקשה מאוד לפרש אותם. שבנינו ותיקפנו ללקוח הוא בעצם היכולת לפרש את המודל.

ניקח דוגמה קיצונית שלא קשורה לסוכן ביטוח. למשל מכונת אוטונומית שיושב בתוכה מודל של למידה עמוקה שמחליט האם עכשיו כשהולך רגל עובר מול המכונת צריך לבלום או לא. גם אם המודל הזה עובד בצורה מצוינת, עדיין יכול להיות שאם שאני לא מבין את המודל עד הסוף ולא מבין מתי הוא טועה – אני פשוט לא אעביר אותו ללקוח.

**שלב 6**

מניסיוני, בכל אחד מחמשת השלבים לעיל, מדען הנתונים בדרך כלל חוזר אחורה, מסדר את הנתונים, בוחר מודל אחר או עושה תיקוף קצת שונה והכל בהתאם לנתונים. משעה שסיימתי את

בלקוח שמשלם פרמיות בסך 7,000 שקל לחודש. נשאלת השאלה, האם אני רוצה שמודל החיזוי שלי ילמד מאותו לקוח? האם החיזויים שלי בעתיד צריכים לקחת בחשבון לקוח כמו אותו לקוח יוצא דופן? אלו שאלות מאוד קשות תחת המטריה הזו של איסוף ועיבוד הנתונים. סוגיה שעולה בהקשר של נתונים היא מהם הנתונים הנחוצים לי לבניית מודל. כי בסופו של דבר הנתונים הגולמיים אינם מוכנים מספיק כדי שאכניס אותם As Is למודל.

**שליבים 3 ו-4**

לאחר שהכנתי את הנתונים לקראת הכנסתם למודל, מגיע השלב השלישי במתודולוגיה, שלב בניית המודל. נעיר רק כי זהו השלב שבעטיו הרבה מאוד אנשים לא מבדילים בין מדע נתונים לבין למידת מכונה. על פי רוב, החלק של בניית המודל הוא השלב שבו נכנסת למידת מכונה וכל המודלים ששומעים עליהם (כגון: רגרסיות, קלסיפיקציות וניתוח אשכולות) נמצאים בשלב הזה.

מאחר ולא ניתן להשתמש במודל חיזוי לפני שעשינו לו תיקוף, הרי שהשלב הרביעי במתודולוגיה הוא שלב תיקוף המודל. בשלב זה אני כמדען מחלק את מאגר הנתונים שבניתי בשלב השני לשני חלקים ביחס של 30-70. באמצעות החלק הגדול (70%) אני מלמד את המודל ובאמצעות החלק הקטן אני בודק את כושר החיזוי של המודל, כאשר פעולה זו נקראת תיקוף. לעולם אין להעביר מודל ללקוח מבלי לתקף אותו קודם לכן.



חצה להבטיח ללקוח שלך הגנה "פרפקט"?  
**בטח שאפשר!**

עכשיו במנורה מבטחים:

**Perfect**

התוכנית הפנסיונית עם ההגנה המקיפה

תמונות



רוצה לשמוע עוד על Perfect? פנה למפקח הרכישה במחוז.