דוד בכר | צילום: עמי ארליך ||
דוד בכר, CFV, CFP, CLU מעריך שווי ומתכנן פיננסי

כלכלן, האם יש לך את הנתונים להיות מדען נתונים?

מדען נתונים מבצע מחקרי מידע מעמיקים כדי להפיק תובנות עסקיות לארגון, מנקה, מטייב ומסדר אתהמידע המשמש למחקרים השונים, מפעיל אלגוריתמים שונים של מידול, כריית מידע ולמידת מכונה על המידע ובונה את תהליכי הכנת המידע והאופטימיזציה של האלגוריתמים השונים. יצאתי לראיין את מדען הנתונים רועי פולניצר על מה צריך על מנת להיות מדען נתונים.

מדען הנתונים רועי פולניצר מכהן כמנכ"ל האיגוד הישראלי למדעני נתונים מקצועיים (PDSIA). רועי בעל תואר M.B.A. במנהל עסקים (עם התמחות בבניית מודלים מתמטיים וסטטיסטיים) ותואר B.A. בכלכלה (עם התמחות בכלים ושיטות לאנליזה ותחקור מידע), שניהם מאוניברסיטת בן-גוריון בנגב ובהצטיינות. רועי למד אקטואריה (בניית מודלים סטטיסטיים מנבאים והסתברותיים לבעיות חיזוי, סיווג, ניתוח אשכולות וזיהוי אנומליות) בכמה מקומות וביניהם בתוכנית ללימודי תעודה באוניברסיטת אריאל בשומרון. בשנה האחרונה רועי ייסד את האיגוד הישראלי למדעני נתונים מקצועיים (PDSIA- Professional Data Scientists' Israel Association) ולכן יצאתי לראיין אותו בנושא.

מה צריך כדי להיות מדען נתונים?

בגדול, מדברים על שלוש מיומנויות שחייבות להיות למדען הנתונים (Data Scientist). המיומנות הראשונה היא הבנה עמוקה ושליטה ברמה גבוהה באלגוריתמים הרלוונטיים של למידת מכונה (Machine Learning), מתמטיקה וסטטיסטיקה. הכוונה היא שעל מדען הנתונים להכיר הרבה אלגוריתמים, להבין את האינטואיציה שלהם, להבין איזה אלגוריתם פחות או יותר מתאים ומתי, מהם היתרונות ומהם החסרונות של כל אלגוריתם.

אוקיי אז המיומנות הראשונה היא להבין ולדעת ליישם אלגוריתמים של למידת מכונה. מהי המיומנות השנייה?

המיומנות השנייה היא היכולת לתכנת בקוד פתוח, כלומר, בשפת Python או בשפת R. אמנם העולם הולך לכיוון של מדע נתונים כשירות אבל כבר ישנם מוצרי מדף עם כל מיני משימות של מדע נתונים.

רגע, אז כבר יש מוצרי מדף שיכולים להחליף את מדען הנתונים?

בגדול כן, אבל לקרוא להם "לא בשלים" זה מחמאה עבורם. לעניות דעתי, ייקח עוד כמה שנים עד שניתן יהיה לבצע חיזוי נטישה באמצעות מוצר מדף, זה עדיין לא נמצא שם ולכן כבר היום מקובל להגיד שמדען צריך להיות מפתח (Developer) בשפת קוד פתוח (בשפת Python, לחילופין בשפת R או לחילופי חילופין בשפת VBA) לכל דבר ועניין, ברמה הנדרשת בתעשייה בכלל ובעולמות ה- Data בפרט

אז המיומנות השנייה היא היכולת לדעת לתכנת בקוד פתוח. מהי המיומנות השלישית?

המיומנות השלישית היא בעצם הבנה עסקית. הכוונה היא לניסיון בעולמי תוכן שונים, יכולת להיכנס, לנתח וללמוד תחומים וענפים שונים, יכולת ללמוד את השפה ואת המאפיינים של ענפים שונים.

למה אתה מתכוון באומרך הבנה עסקית, תוכל להסביר בבקשה?

כן. יש שלב שבמדע נתונים נקרא "הנדסת מאפיינים" (Features Engineering, כאשר המאפיינים הם המשתנים המסבירים) שבו אני לוקח את הנתונים הגולמיים של הלקוח והופך אותם למידע מעניין יותר או רלבנטי יותר. לדוגמא, אם אתה "משתמש" (User) במשחק אפליקציה מסוים ויש לך הרבה כניסות ופעולות במשחק, אז אי אפשר להכניס למודל של למידת מכונה את כל הפעולות האלה ולכן עליי כמדען הנתונים להשתמש את בהבנה העסקית שלי ולנסות לחשוב איך אני מתרגם את אוסף הפעולות האדיר הזה שלך למאפיינים מאוד מאוד ספציפיים. כמו למשל, כמה פעמים ביום אתה כמשתמש נכנס למשחק? כמה פעמים בחודש אתה נכנס למשחק? כמה זמן עובר בין כל 2 כניסות סמוכות שלך? כמה זמן נמשכת כל כניסה שלך? אלו סוגי פעולות אתה עושה באפליקציה? כל הדברים הללו נועדו, בעיניי לפחות, לאפיין אותך כמשתמש באמצעות מאפיינים קצת יותר מעניינים, קצת יותר אגרגטיביים שתופסים את המהות שלך כמשתמש. וזה למעשה חלק מהבנה עסקית.

זה מעניין. ותגיד, אין מקרים שבהם מדען הנתונים מגיע ל"שוקת שבורה" בגלל שאין לו שפה משותפת עם הלקוח?

בהחלט. תיקח למשל פרויקט של מדע נתונים באיזשהי חברת שמייצרת תקשורת בתוך רכבים, כאשר אני כמדען הנתונים לא יודע דבר וחצי דבר בפרוטוקולים של תקשורת של רכבים. אבל מה לעשות כמדען נתונים אין לי הרבה ברירה, אז או שאני לומד את התחום כחלק מההכשרה שאני מעביר את עצמי לפני תחילת כל פרויקט (בפרויקטים מורכבים אורך ההכשרה יכול לעלות אפילו על 60 שעות) על מנת שתהיה לי שפה משותפת עם הלקוח או שאני אומר תודה, לא מתאים לי ופשוט לא לוקח את הפרויקט. אבל בשורה התחתונה על מדען הנתונים ללמוד את עולם התוכן הספציפי של הלקוח על מנת שהוא יוכל לייצר עבורו מאפיינים (Features) חזקים יותר.

אז למעשה הידע הכלכלי והעסקי המקצועי שלך כמדען נתונים שהוא גם אקטואר, גם מנהל סיכונים פיננסיים וגם מעריך שווי הוא יתרון.

בדיוק כך. חשוב להבין ששום קורס של מדע נתונים או למידת מכונה לא יכול להקנות לסטודנט הבנה עסקית, אינטואיציה עסקית, ומעל לכול סקרנות ויצירתיות ולא יכול להעניק לו את הידע הנדרש לכל ענף ולכל עולם תוכן כמו גם את היכולת לשים לב לפרטים ולקשרים שונים ומפתיעים. לדוגמא, אם למדען נתונים מסוים יש נתונים על מחירי דירות במינכן-גרמניה ונניח שאותו מדען נתונים הוא לא שמאי מקרקעין ואפילו לא בעל תואר ראשון בכלכלה אלא נניח בעל תואר שני במתמטיקה ולכן הוא לא יודע לעשות העשרת נתונים (Data Enrichment). לאמור- הוא לא יודע מה חשוב ומה לא והוא לא יודע איך השוק הזה עובד, אז אותו מדען נתונים יבוא בתמימות ויפעיל כל מיני אלגוריתמים כאלה ואחרים. בסוף היום, מדען הנתונים שיביא איתו ידע מקצועי בעולמות תוכן שונים ובכלל זה בעולם התוכן הספציפי של הלקוח, הוא זה שיהיה לו יתרון ברור על כל מדען נתונים אחר שאין לו את הידע המקצועי הזה.

אני הרבה זמן רוצה לשאול אותך, מה הקשר בין מדע נתונים לבין סטטיסטיקה?

שאלה טובה. להרבה מאוד אנשים יש חוסר היכרות עם עולם מדע הנתונים. אם נסתכל על מתודולוגיית 6 השלבים של CRISP-DM (הבנת הבעיה, איסוף ועיבוד הנתונים, בניית המודל, תיקוף המודל, בדיקת הגיוניות המודל ומסירת המודל), אז הסטטיסטיקה למעשה נכנסת רק בחלק מהשלבים: בשלב איסוף ועיבוד הנתונים ובשלב בניית המודל. כך למשל בשלב איסוף ועיבוד הנתונים אני כמדען נתונים יכול לעשות שימוש בסטטיסטיקה על מנת להגיד שמשתנה מסוים הוא לא משמעותי, לחילופין שהמשתנה הזה והזה הוא לא אינפורמטיבי דיו, או לחילופי חילופין ששני המשתנים האלה אומרים בדיוק את אותו הדבר, ואז אני יכול לוותר על אחד מהם. בשלב בניית המודל אני יכול לעשות שימוש בסטטיסטיקה על מנת לכוון את האלגוריתמים שלי באמצעות כל מיני הנחות סטטיסטיות, לחילופין לכוון את היעד (Target, המשתנה המוסבר) שלי כך שיתפלג בצורה מסוימת.

אז למעשה סטטיסטיקה היא פשוט עוד כלי בארגז הכלים של מדען הנתונים?

בדיוק כך. הסטטיסטיקה היא כלי נוסף בארגז הכלים של מדען הנתונים, הא ותו לא. אמנם זהו כלי חשוב וזה כלי שהרבה מאוד פעמים מבלבלים בינו לבין מדע נתונים, אבל זה עדיין לא חזות הכל. כך למשל יכול להיות סטטיסטיקאי או מתמטיקאי או מהנדס תוכנה או איש מדעי המחשב מדהים שלא יצליח להיות מדען נתונים טוב כי אין לו הבנה עסקית טובה או בכלל, כלומר, שהוא לא מתחבר לכלכלה ולמנהל עסקים כי הוא בסך הכל יודע לתכנת מאוד מאוד טוב אבל מתקשה להיכנס כל חודש לענף או לתחום חדש וללמוד אותו על בוריו. מאידך, יכול להיות כלכלן או רואה חשבון שלא יצליח להיות מדען נתונים טוב כי אין לו יכולת תכנות טובה, כלומר, שהוא לא מתחבר למדעי המחשב ולהנדסת תוכנה כי הוא בסך הכל יודע לאסוף נתונים ולנתח ענפים וחברות מאוד מאוד טוב אבל מתקשה לפתח שורות קוד, לא מסתדר עם וקטורים, מטריצות, תנאים, לולאות, פונקציות, מחרוזות, וכו'. על כן, הסטטיסטיקה היא עוד איזשהו כלי בארגז הכלים ואמנם כבודה

במקומה מונח אבל הא לא יותר מכלי מרכזי.

ומהו דינה של המתמטיקה עבור מדען הנתונים?

לעניות דעתי, המתמטיקה היא מלכת המדעים. אבל בקונטקסט של כישורים של מדען נתונים, הרי שדינה של הסטטיסטיקה הוא הדין גם לגבי חשבון דיפרנציאלי ואינטגרלי, אלגברה לינארית, טופולוגיה, אנליזה אופטימיזציה ועוד הרבה מאוד כלים מתמטיים שמשתמשים בהם במדע הנתונים. חשוב להבין, מדע הנתונים שומר על איזשהו ריחוק מהתיאוריה והוא די אינטואיטיבי ופחות מתמטי. כמובן שכל אחד לוקח את מדע הנתונים לכל מיני מקומות. אחד לוקח את מדע הנתונים יותר לכיוונים של השגת הנתונים, אינטגרציה של הנתונים ממספר מקורות, עבודה עם מאגרי נתונים גדולים (Big Data) ועיבוד נתונים לא מובנה (Unstructured). אחר לוקח את מדע הנתונים יותר לכיוונים של חקירת הנתונים, תכנות, ניתוח סטטיסטי ובניית מודל לתחקור. שלישי לוקח את זה יותר לכיוונים של ניתוח אנליטי של הנתונים, חיזוי, כריית מידע, אופטימיזציה, עיבוד מידע טקסטואלי ואנליזה של נתונים גדולים. ורביעי לוקח את זה יותר לכיוונים של הצגת הנתונים, הצגת תוצרי התחקור ובניתת כלי ויזואליזציה שונים.

שאלה היפותטית משהו, מה הקשר בין מדע הנתונים לבין מאגרי נתונים גדולים (Big Data)?

חשוב לי מאוד להגיד שאין שום קשר.

איך זה יכול להיות הרי מדע נתונים ומאגרי נתונים גדולים באים ביחד.

אז חשוב לי להגיד את זה בצורה מאוד מאוד ברורה: אין קשר בין מדע נתונים לבין מאגרי נתונים גדולים. בשני הראיונות האחרונים הסברתי באריכות מהו מדע נתונים. לגבי מאגרי נתונים גדולים זה כבר עניין תשתיתי פרופר. כלומר, מאגרי נתונים גדולים נותנים לך אמנם את האפשרות לעבוד עם הרבה נתונים, ETL עם הרבה נתונים, שמירה עם הרבה נתונים, העברה של הרבה נתונים, זרימה של הרבה נתונים. כל הדבר הזה דורש תשתיות ולצורך היכרות עם תשתיות יש היום קורסים של מאגרי נתונים גדולים. אבל למדעני נתונים בצרכים שלהם, בנגזרת שלהם, אין עניין במאגרי נתונים גדולים. לאמור- הידע שמדען הנתונים "מביא איתו לשולחן" לא קשור לטכנולוגיה שבה "יושבים" הנתונים. רוצה לומר- שכל האלגוריתמים האלה של למידת מכונה, כל הנדסת המאפיינים הזאת וכל הסטטיסטיקה הזו – כולם ישימים ונכונים לנתונים ללא שום קשר לגודלם של הנתונים. יחד עם זאת, כיום מאגרי נתונים גדולים זו טכנולוגיה מאוד מאוד חשובה שקיימת בשוק ולכן המקום שבו מאגרי נתונים גדולים פוגשים את המדע הנתונים זה Spark. אבל אני לא אכנס לזה עכשיו.

אז מהן היום דרישות התפקיד ממדען נתונים באירגונים גדולים בתחומים של אנליטיקה פיננסית או שיווקית?

דרישה אחת ש"מסתובבת לי בראש" אני ומדבר איתך היא שימוש מתקדם בדאטה (Big Data) באמצעות בניית מודלים, פיתוח ושימוש באלגוריתמים של למידת מכונה וניתוח תהליכים לצורך זיהוי כיוונים ומגמות במגוון תחומים לרוחב הארגון.

אוקיי, עבודה עם מאגרי נתונים גדולים תוך פיתוח ושימוש באלגוריתמים של למידת מכונה על מנת לזהות כיוונים ומגמות בעולמות תוכן שונים. מה עוד?

דרישה שנייה, יכולה להיות אחריות על זיהוי אתגרים עסקיים שבהם הנתונים יכולים להוות גורם מכריע בשיפור קבלת ההחלטות. דרישה שלישית, עשויה להיות אחריות על איתור ואיסוף מקורות מידע פנים ארגוניים וחיצוניים, הגדרה ואיפיון של שימושי המידע בארגון. דרישה רביעית, יכולה להיות הבניית מאגרי נתונים שאינם מובנים ונמצאים במסגרת הארגון, אפיון והגדרת הצגת המידע ותוצריו לדרג מקבלי ההחלטות בארגון. דרישה חמישית, יכולה להיות פיתוח כלים, מודלים, תהליכים ומערכות לרוחב הארגון בתחום האנליזה.

לסיכום, איך אתה רואה את תפקידו של מדען הנתונים?

תפקידו של מדען הנתונים הוא לבצע מחקרי מידע מעמיקים בכדי להפיק תובנות עסקיות לארגון, לנקות, לטייב ולסדר את המידע המשמש למחקרים השונים, להפעיל אלגוריתמים שונים של מידול, כריית מידע ולמידת מכונה על המידע, ולבנות את תהליכי הכנת המידע והאופטימיזציה של האלגוריתמים השונים. ככה אני מגדיר מדען נתונים.

מדען נתונים הוא אחד שגם טוב יותר בסטטיסטיקה ואקונומטריקה מכל איש מדעי המחשב או מהנדס תוכנה וגם טוב יותר בהנדסת תוכנה מכל סטטיסטיקאי או כלכלן.