

מנהל סיכונים פיננסיים, על מדע נתונים כבר שמעת?

מדע נתונים (Data Science) הוא תחום שבו מסיקים מסקנות מתוך נתונים. יצאתי לראיין את רועי פולניצר על התחום המרתק הזה שנקרא מדע נתונים

השוני נעוץ במתודולוגיה המכונה מתודולוגיית 6 השלבים של CRISP-DM (תהליך סטנדרטי חוצה-ענפים לכריית נתונים), לפיה השלב הראשון בכל פרויקט מדע נתונים זה בעצם הבנת הבעיה. על פניו זה נשמע מצחיק, מהי כבר יכולה להיות אותה בעיה גדולה? אבל מסתבר שמדובר בבעיה די מורכבת מאחר ובסופו של דבר המטרה שלנו היא לתרגם בעיה עסקית לבעיה מתמטית, בעיה שאותה אנו הולכים לפתור בכלים מתמטיים.



מדען הנתונים רועי פולניצר משמש 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 מחזיק בכמה הסמכות מקצועיות רלוונטיות למדע נתונים ולמידת מכונה, ביניהן הסמכת "אקטואר מלא" (Fellow) מטעם לשכת מעריכי השווי והאקטוארים הפיננסיים בישראל (IAVFA- Israel Association of Valuators and Financial Actuaries) המעידה על כך שהמחזיק בה בקיא בפיתוח, יישום ותיקוף מודלים סטטיסטיים ואלגוריתמים מתמטיים כגון RF, GLM ו-NN לקביעת פרמיות בביטוח כללי). בנוסף, רועי חבר באיגוד הבינלאומי למנהלי סיכונים מקצועיים (PRMIA- Professional Risk Managers' International Association), מה שמעיד עליו כבקיא במחקר, פיתוח, כתיבה ומימוש אלגוריתמים של בינה מלאכותית על כמותיות גדולות של מידע. בשנה האחרונה רועי ייסד את האיגוד הישראלי למדעני נתונים מקצועיים (PDSIA- Professional Data Scientists' Israel Association) ולכן יצאתי לראיין אותו בנושא.

מה זה מדע נתונים?

אם למשל מישהו שואל על בעיית נטישת לקוחות או רוצה לדעת לזהות מתי הלקוחות שלו עומדים לנטוש אותו, זוהי אמנם שאלה בעברית והיא נשמעת שאלה מעניינת, אבל כשאני כמדען נתונים רוצה להתחיל לתרגם את זה למספרים אז מתחילות להעלות לי כל מיני שאלות אחרות, כמו למשל: מה זה לקוח? האם לקוח זה מישהו שנמצא אצלך לפחות שנה? האם לקוח זה מישהו שמשלם לך כל חודש? מי זה לקוח אצלך? וזה מאוד תלוי בעסק. שאלה אחרת היא מה זה נטישה? האם נטישה זה בנאדם שכבר לא נמצא ברשימת הלקוחות שלך? או האם נטישה זה בנאדם שלא תקשר איתך שנה? שאלה נוספת היא מה זה בכלל חיזוי נטישה? האם אני צריך לדעת חודש מראש? שבועיים מראש? יום מראש? כל הדברים הללו אלו דברים שצריך להגדיר אותם היטב וגם להגדיר את המדדים שלהם. בסופו של דבר אנחנו נבנה

בגדול, מדע נתונים (Data Science) הוא תחום שבו מסיקים מסקנות מתוך נתונים. הבעיה שברגע שמסתפקים בהגדרה הצרה הזו אז למעשה Data אנליסטים, אנשי ETL (הוצאת נתונים מקבצי המקור, ביצוע טרנספורמציה לנתונים וטעינת הנתונים שעברו טרנספורמציה) ומתכנני BI (בינה עסקית) אומרים שהם עושים את אותו הדבר והאמת שהם צודקים. לאמור- העבודה של הרבה מאוד אנשי נתונים היום היא למעשה מדע נתונים מכל מיני היבטים.

האם תוכל להסביר במה Data Science קצת שונה או קצת יותר רחב מהתחומים האחרים שצינית: ETL ו-BI?

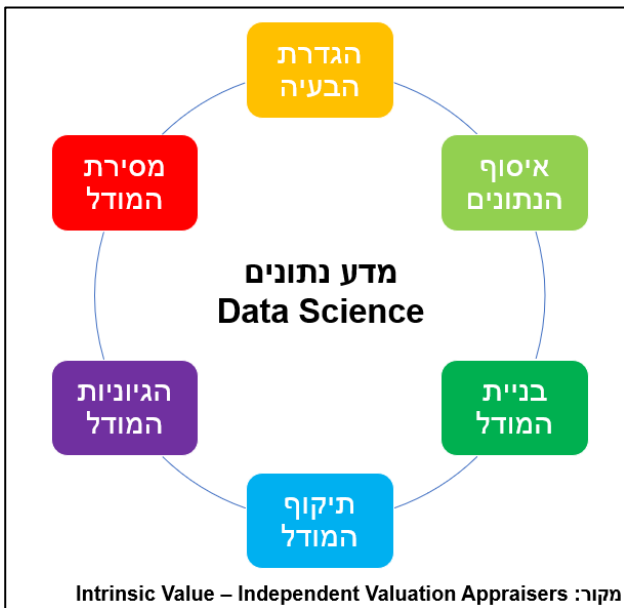
תיקוף. חשוב לציין שגם אם המודל שלי עבר תיקוף, אני עדיין לא אעביר אותו ללקוח על מנת שזה יעשה בו שימוש לקבלת החלטות בזמן אמת. אם דיברתי מקודם על עצמי כלקוח באיזשהו משחק סולוארי, אז כתוצאה מאיזשהו מודל, החברה של המשחק יכולה להחליט עכשיו לשלוח לי איזשהי הודעה, כי המודל אומר לה שאם עכשיו אני אקבל את ההודעה הזו, אז אולי אני אכנס שוב למשחק או לחילופין אולי אני אקנה משהו. זה מה שנקרא להעביר את המודל ללקוח.

אז מה עושים לפני שמעבירים את המודל ללקוח?

שלב אחד לפני שאני מעביר את המודל ללקוח, אני כמדען נתונים רוצה להבין האם המודל שלי הוא הגיוני. לאמור- האם מה שהמודל אומר זה הגיוני? האם יש לזה משמעות בכלכלה?

אז השלב החמישי הוא שלב בדיקת הגיוניות המודל?

אכן כן. אני כמדען נתונים מקדיש זמן לא מבוטל בלנסות ולהבין מה המודל אומר לי. כך למשל, מודל רגרסיה לינארית מספר לי סיפור בצורה מסוימת. מודל עץ החלטה (Decision Tree), ששואל שאלות על המודל, נותן לי אינדיקציות אחרות ויש כמובן מודלים של למידה עמוקה (Deep Learning) שההסבר שלהם או שהפרשנות שלהם היא מאוד דלה. לאמור- קשה לי מאוד להבין מה מודלים של למידה עמוקה אומרים לי ולכן ה"חם" להעברת המודל ללקוח זה בעצם שלב בדיקת הגיוניות המודל. אם ניקח דוגמה קיצונית, אז למשל מכונית אוטונומית שיושב בתוכה מודל של למידה עמוקה שמחליט שאם עכשיו שכשוהלך רגל עובר מול על המכונית לבלום. אז גם אם המודל הזה עובד בצורה מצויינת, אז עדיין יכול להיות שמאחר ואני לא מבין את המודל עד הסוף ולא מבין מתי המודל הזה טועה אז אני לא אעביר אותו ללקוח.



נניח שעברנו את כל 5 השלבים האלה, ואגב אני מתאר לעצמי שבכל שלב ניתן לחזור אחורה ולסדר קצת את הנתונים, לחילופין לבנות מודל קצת אחר או לחילופין חילופין לבצע תיקוף קצת שונה והכל בהתאם לנתונים, מהו השלב השישי והאחרון במתודולוגיה?

ברגע שסיימתי את חמשת השלבים הללו, אני מגיע בשעה טובה לשלב השישי, הלא הוא שלב מסירת המודל ללקוח. בשלב הזה אני כמדען

איזשהו מודל לחיזוי נטישה וצריך לדעת האם המודל הזה עובד או לא עובד ואיך מודדים את זה.

אז השלב הראשון של מדע נתונים הוא הגדרת הבעיה. מהו השלב השני?

אחרי השלב הראשון של הגדרת הבעיה מגיע השלב השני של איסוף ועיבוד הנתונים. לא בכדי נהוג לומר שמדע נתונים זה 90% עבודה עם הנתונים. בשלב זה, אנו אוספים את הנתונים, "מנקים" אותם, מכינים אותם, מעבדים אותם. זה אומר שאם יש לי מידע חסר אז עליי כמדען נתונים להחליט האם אני משלים אותו או זורק אותו. מה קורה אם ישנה עמודה מסוימת בנתונים שחסרה הרבה נתונים? האם אפשר לוותר על העמודה הזאת? מצד שני אולי זו עמודה מאוד חשובה שאני לא רוצה לוותר עליה? מה קורה אם יש לי כל מיני Outliers (חריגים), כלומר, דוגמאות מאוד קיצוניות. נניח שאני מנסה להבין משהו על הלקוחות של חברה מסוימת ורב הלקוחות שלה קונים ב- 100 ש"ח לחודש או 200 ש"ח לחודש ופתאום אני נתקל בלקוח שקונה ב- 7,000 ש"ח לחודש. נשאלת השאלה, האם אני רוצה שהמודל שלי ילמד מאותו בנאדם? האם החיזויים שלי בעתיד צריכים לקחת בחשבון לקוח כמו אותו לקוח יוצא דופן שקנה ב- 7,000 ש"ח לחודש?

אלו שאלות מאוד מאוד קשות תחת המטריה הזו של איסוף ועיבוד הנתונים.

ישנה עוד נושא שעולה בהקשר של איסוף ועיבוד הנתונים והיא מהם הנתונים שצריכים בסופו של דבר להיכנס למודל. כי בסופו של דבר הנתונים הגולמיים אינם "מוכנים" בצורה מספיק טובה.

למה אתה מתכוון באומרך לא מספיק טובה?

נגיד שאני רוצה לאפיין לקוח מסוים באמצעות כל הטרנזקציות שלו. לדוגמה, ניקח חברה למשחקי אפליקציה שרוצה לאפיין אותך כלקוח שלה. אבל הנתונים הגולמיים שקשורים אליך הם מאוד מורכבים, מדובר על הרבה מאוד כניסות שלך לאפליקציה, הרבה מאוד פעולות שלך באפליקציה. רוצה לומר- החברה צריכה איכשהו להפוך את הנתונים הללו ממשוהו שמאוד פרטני למישהו מאוד אגרסיבי שיאפיין אותך בצורה טובה. למשל, כמה פעמים ביום אתה נכנס לאפליקציה? כמה זמן נמשכת כל שהייה שלך באפליקציה? ודברים מהסוג הזה. אז ההכנה הזו טרום עיבוד הנתונים היא למעשה תהליך מאוד מורכב.

ואחרי שאנחנו שטיפלנו בנתונים והכנו אותם בצורה מספיק טובה לקראת הכנסתם למודל, מהו השלב השלישי?

השלב השלישי הוא שלב בניית המודל. זהו החלק שבגללו הרבה מאוד אנשים לא מבדילים בין מדע נתונים לבין למידת מכונה (Machine Learning). על פי רוב, החלק של המודלים הוא תחת המטריה של למידת מכונה וכל המודלים והאלגוריתמים ששומעים עליהם, כגון: רגרסיות, קלסיפיקציות ו-k-Nearest Neighbors (אלגוריתם השכן הקרוב ביותר או k-NN) וכו' נמצאים בשלב הזה שבו אני כמדען נתונים בונה את מודל החיזוי/הסיווג/ניתוח האשכולות. הערה אינפורמטיבית: לא עושים שימוש במודל לפני שתקפנו אותו.

אז אחרי שבנינו את המודל, מה השלב הרביעי?

השלב הרביעי הוא שלב תיקוף המודל. מי שקרא חלק מהמאמרים שלי בנושא למידת מכונה ודאי מכיר את הקונספטים של סט אימון וסט בדיקה. למעשה אני כמדען נתונים מחלק את הנתונים שלי לשני חלקים. באמצעות חלק אחד (70% מהנתונים) אני מלמד את המודל ובעזרת החלק האחר (30% הנותרים) אני בודק את המודל וזה נקרא תיקוף. חשוב להבין שאני כמדען נתונים לעולם לא אעביר מודל ללקוח שלי אם הוא לא עבר איזשהו

נתונים יכול כבר למסור את המודל ללקוח.

הרבה אנשים שחושבים על מדע נתונים, לרוב חושבים על שלב בניית המודל ובעבר אני זוכר שהיו קוראים לתחום הזה מידול סטטיסטי או אקונומטריקה, מה דעתך על כך?

זוהי טעות גמורה, הואיל וכיום מדע הנתונים תופס את כל התהליך שתיארתי לעיל כאיזשהו תרשים זרימה שלם שהוא בעצם מדע הנתונים. זה כמובן משתנה מחברה לחברה לפי כוח האדם שלה, לפי היעדים שלה ולפי האתגרים הטכנולוגיים שלה.

אז הסטטיסטיקה למעשה נכנסת רק בשלבים של איסוף הנתונים ובניית המודל?

לחלוטין. למשל בשלב איסוף הנתונים אני יכול להשתמש בסטטיסטיקה כדי להגיד שמאפיין (Feature, משתנה מסביר) מסוים הוא לא משמעותי או לא אינפורמטיבי או לחילופין ששני המשתנים האלה אומרים את אותו הדבר ואז אני יכול לוותר על אחד מהם. לאמור- כל מיני שיקולים כאלה וסטטיסטיקות מסוגים שונים. בשלב בניית המודל אני יכול לכוון את המודלים שלי באמצעות כל מיני הנחות סטטיסטיות או לחילופין לכוון את היעד (Target, המשתנה המוסבר) שלי להיות עם איזושהי התפלגות מסוימת.

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

לגמרי. הסטטיסטיקה היא אמנם כלי מאוד חשוב שהרבה מאוד פעמים מבלבלים בינו לבין מדע נתונים אבל היא איננה חזות הכל ובטח לא הכלי המרכזי. הדבר החשוב ביותר הוא הבנה עסקית ולכן לדעתי מדעני נתונים טובים יותר הם אלו שדווקא באים מתחום היעוץ הכלכלי ולא סטטיסטיקאים וממתטיקאים שנעדרי כל יכולת לנתח עסקים וענפים וללמוד עולמות תוכן.

איזה ניסיון מקצועי מביא אותך לתחום ה- Data Science?

ראשית, יש לי ניסיון רב בפיתוח ויישום מודלים כמותיים, במיוחד בתחום ניהול הסיכונים. לאמור- אני כיועץ מסייע ללקוחותיי לפתח וליישם מודלים מתקדמים הדורשים הבנה עמוקה בתהליכים סטוכסטיים, ידע בשיטות נומריות ויכולות פיתוח תוכניות בשפות ניתוח נתונים, כגון: Python ו-R. אל תשכח שמזה 12 שנים שאני נותן ייעוץ בניית תוחמים כמותיים מתקדמים בתחומים של הנדסה פיננסית, יישום מודל מונטה-קרלו, תהליכים סטוכסטיים ופתרון בעיות כמותיות באמצעות שיטות נומריות מתקדמות.

איזה עוד ניסיון?

יש לי גם ניסיון רב בבניית פתרונות ממוחשבים מבוססי אקסל. כידוע לך תוכנת אקסל נמצאת בשימוש של ארגונים רבים, אולם בדרך כלל יכולותיה מנוצלות באופן חלקי מאוד. שימוש בכלים מתקדמים שמציעה תוכנת האקסל מאפשרת יצירת קבצים, ההופכים את העבודה לנוחה ויעילה יותר. קבצים אלו מאפשרים ביצוע משימות שהיו נראות בלתי אפשריות ברמת אקסל, בעזרת שימוש בפונקציות מתקדמות או בניית פונקציות ייחודיות לארגון בעזרת VBA, שפת תכנות המובנית באקסל (VBA- Visual Basic Applications for). למעשה אני בונה פתרונות ממוחשבים לארגונים באמצעות קבצי אקסל המותאמים לצרכיהם, הכוללים תכונות מתקדמות של אקסל ו-VBA. הקבצים נבנים לאחר לימוד צרכי המשתמשים ואפיון משותף של דרישות הארגון והמשתמשים תוך יישום נושאים שונים במימון בתוכנת אקסל, כגון: מודל Black & Scholes, מודל בינומי C-R-R, מודל Monte Carlo, מודל ה-VaR, גלאי 2 ונושאים נוספים לפי דרישה.

הזכרת תיקוף המודלים, האם יש לך ניסיון בתיקוף מודלים?

בהחלט. אבל בכל זאת כמה מילים על תיקוף מודלים. בעידן שבו הולך וגובר שימוש במודלים כמותיים מורכבים, חשיבותו של תהליך תיקוף המודלים איננה מוטלת בספק. הדבר בא לידי ביטוי ברגולציה בארץ ובעולם. כך לדוגמה, כבר לפני 11 שנים הוציא בנק ישראל מכתב שבו הוא מורה לבנקים להקים פונקציה לתיקוף המודלים המשמשים אותם בחישובי שווי הוגן של מכשירים פיננסיים כמו גם את מודלי תחזיות בתחום ניהול הסיכונים. על פי באזל II (הרגולציה הבינלאומית לניהול סיכונים בבנקאות) תהליך התיקוף אמור לכלול גם מודלים לסיכונים שוק וסיכונים אשראי הנמצאים בשימוש הבנק. הוא הדין לגבי סולבנסי II (הרגולציה הבינלאומית לניהול סיכונים בביטוח). אני יכול לומר לך שבבואי לבצע תיקוף מודל, אני מתייחס לשלושת המרכיבים הבאים: (1) ההנחות ונתוני קלט של המודל; (2) התיאוריה שמאחורי המודל ואופן יישומה; (3) הדיווח על סמך המודל. כמובן שבעת בחינת המרכיבים הללו אני סוקר את המודל, משווה בין תוצאותיו לבין תוצאותיהם של מודלים אחרים כמו גם בין תחזיותיו לבין התוצאות בפועל.

אז אני מבין שיש לך ניסיון בתיקוף?

בוודאי. אני רק רוצה לציין שכבר למעלה מעשור שנים מתמודדים גופים רבים, הכוללים חברות ביטוח, בתי השקעות, קרנות פנסיה וקופות גמל, עם הטמעת מערכות לניהול סיכונים. מערכות אלו מספקות, בין היתר, תוצאות חישוב של מדדי סיכונים כגון Value-at-Risk (VaR). אני לקחתי חלק בשלבים המאוחרים של תהליכי ההטמעה הללו שכללו טיפול במכשירים לא סטנדרטיים (מכשירים אקזוטיים ולא סטנדרטיים, כאשר תמחור מכשירים אלו כרוך בניית מודלים ייחודיים, המתבססים על טכניקות נומריות מתקדמות) ותיקוף התוצאות.

אני שמעתי ששימוש במערכות לניהול סיכונים מלווה בעצמו בסיכון מודל משמעותי?

אמת. חישובי ניהול סיכונים מבוצעים על תיקים הכוללים מאות סוגים של מכשירים פיננסיים, והתוצאות המתקבלות אינן בהכרח אינטואיטיביות. במקרים רבים, מערכות לניהול סיכונים הינן מעין "קופסאות שחורות", ומשתמשים רבים המפעילים אותן מודאגים (ובדיון) מנכונות התוצאות המתקבלות מאותן מערכות. על מנת להפיג חשש זה מומלץ לבצע תהליך של תיקוף התוצאות.

אלו שיטות קיימות לתיקוף?

ככלל, קיימות מספר של שיטות לתיקוף תוצאות מטבע הדברים, לכל שיטה יש יתרונות וחסרונות ויש לנקוט בה בהתאם למידת ישימותה למקרה המתאים ו/או לסיטואציית ההערכה ומטרתה. השיטה הראשונה לתיקוף תוצאות היא ביצוע חישובים מקורבים. כך לדוגמה, בתיקים החשופים לסיכון שער ריבית ניתן לבדוק את סבירות התוצאות מתוך נתוני סטיות תקן ומח"מ (כאשר חשוב לא להתבלבל בין Price Volatility ל- Yield Volatility). השיטה השנייה הינה ביצוע חישובים בלתי תלויים בתוכנת אקסל. השיטה השלישית הינה ביצוע חישובים בלתי תלויים בתוכנות אחרות לניהול סיכונים.

אז תהליך התיקוף הנו קריטי.

מאוד. לעניות דעתי מומלץ לבצע תיקוף לפני שלב פרסום התוצאות המופקות במערכות לניהול סיכונים. אני למשל מבצע תהליך שכזה גם עבור חברות אשר הטמיעו את המערכת לניהול סיכונים או נמצאות בשלבי סיום. ובדומה לתחומי הייעוץ האחרים שלי.

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