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

בעל חברה, מודל לחיזוי נטישת לקוחות כבר הזמנת?

בניית מודל לחיזוי נטישת לקוחות תלויה בקיומם של נתונים. על לקוחות להבין שקיימת גם אפשרות שלא ניתן יהיה לבנות מודל נטישה עם הנתונים הקיימים.

רועי פולניצר הצליח להיות מעריך שווי, אקטואר ומנהל סיכונים פיננסיים מצליח ולסיים בהצטיינות לימודי דיפלומה בניהול סיכונים פיננסיים כמו גם ללמוד לימודי תעודה באקטואריה לפני שפצח בקריירה של מדען נתונים. מאז הספיק כבר רועי לעשות שימוש בדאטה (Big Data) לצורך בניית מודלים, פיתוח ושימוש באלגוריתמים של Machine Learning לצורך חיזוי, סיווג וניתוח אשכולות וניתוח כמו גם לתקף מודלים בעולמות תוכן, כגון: הערכות שווי, ניהול סיכונים, אקטואריה והנדסה פיננסית וכו'. רועי מתהדר בשלל תארים. מהאקדמיה קיבל תואר שני במנהל עסקים ותואר ראשון בכלכלה, שניהם בהצטיינות ובמבחנים הבינלאומיים להסמכה בניהול סיכונים פיננסיים (FRM) מטעם האיגוד העולמי למומחי סיכונים (GRAP) בניו-ג'רזי בארה"ב רועי התברג בין 160 מנהלי הסיכונים הפיננסיים הטובים ביותר בעולם מתוך 16,000 סטטיסטיקאים ואקונומטריקאיים מ- 90 מדינות. וכל זה לא כולל כמובן את ההסמכות המקצועיות שלו כ"אקטואר מלא" (Fellow) מטעם לשכת מעריכי השווי והאקטוארים הפיננסיים בישראל (IAVFA) וכמומחה לניהול סיכונים (CRM) מטעם האיגוד הישראלי למנהלי סיכונים (IARM). בשנה האחרונה רועי התחיל לכתוב על תחום מדע הנתונים ולמידת המכונה ולכן יצאתי לראיין אותו בנושא.

נטישת לקוחות (Customer Churn) מתרחשת כאשר לקוחות או מנויים מפסיקים לעשות עסקים עם חברה כלשהי או שירות מסוים. נטישה לקוחות, המכונה גם אזילת לקוחות (Customer Attrition), הינה מדד קריטי מכיוון שהרבה יותר זול לשמר לקוחות קיימים מאשר לרכוש לקוחות חדשים - להרוויח עסקים מלקוחות חדשים פירושו עבודה כל הדרך דרך "משפך המכירות", דהיינו, שימוש בשיווק ובמשאבי המכירות של החברה לאורך כל התהליך. מאידך, שימור לקוחות (Customer Retention) הינו על פי רוב חסכוני יותר מאחר שהחברה כבר הרוויחה את האמון והנאמנות של הלקוחות הקיימים.מדעני נתונים עוזרים לחברות לפתח מודלים מנבאי נטישת לקוחות וקיימים אלגוריתמים רבים שבאמצעותם ניתן לפתח מודלים לניבוי נטישת לקוחות.

נניח וקיבלת הצעה מבעל חברה שיש לו 10,000 לקוחות לבנות לו מודל נטישת לקוחות על טהרת מדע נתונים, מה קורה מכאן?

אני חייב לשים על השולחן את האפשרות שלא ניתן לבנות מודל נטישה עם הנתונים הקיימים. אתה חייב להבין ש- 10,000 לקוחות זה מעט מדי לקוחות. בנוסף, צריך להבין מה זה אומר נטישה. לפעמים נטישה זה לא משהו מובן מאליו כמו שאנחנו חושבים על זה בעברית 1 או 0. צריך להגדיר קבועי זמנים, יצירת קשר וכיוצא באלה פרמטרים. זה המון המון עבודת הכנה והמדדים הסופיים הם לא ברורים.

כשלקוח שואל אותי כיצד מזהים מתי לקוח מסוים עומד לנטוש אותו או מתי זה גם יקרה, אז 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 אני יבנה איזהשהו מודל לחיזוי נטישה וצריך לדעת האם המודל הזה עובד או לא עובד ואיך מודדים את זה. לשאלתך אני לא חושב ש- 10,000 לקוחות זה מספיק נתונים.

וננייח שלי כבעל חברה יש מאגר אינסופי של נתונים?

זה כבר סיפור אחר. חשוב להסביר שזה לא משנה כמה עמודות יש לך לתת לי, מה ששמשנה לי זה כמה שורות יש לך לתת לי. בסופו של דבר מספר הלקוחות שלך, בעבר ובהווה כאחד, זה הנתונים שאיתם אני כמדען נתונים עובד. למעשה כל שורה בנתונים שלך מאפיינת לקוח אחד. ואם יש לך 10,000 לקוחות זה לא כמו שיש לך מיליון לקוחות. זה לא אותו דבר.

במה זה שונה?

האפשרות לעשות סגמנטציה, כלומר, פירוק של הנתונים לקבוצות הגיוניות היא פחות טובה, האפשרות לנטרל "רעשים" היא פחות טובה. כל המודלים שאמורים לעזור לי כמדען נתונים לנקות את הנתונים שלך יעבדו פחות טוב. וכך אני אשאר לבסוף עם מעט מאוד נתונים דבר שיקשה עליי להסיק נתונים מתוך אותם נותנים. לכן נקודת הפתיחה היא כמות הנתונים שיש לך לתת לי.

אבל אתה אמרת לי בראיון אחר שמדע נתונים ומאגרי נתונים גדולים (Big Data) לא חייבים ללכת ביחד.

אמת. אבל כידוע לך החוזק של המודלים נמדד לפי גודל המדגם שעל בסיסו הם פותחו. כשיש לי כמדען נתונים מעט תצפיות אז המודל שאני אבנה לא צפוי להכליל בצורה טובה נתונים שלא שימשו לפיתוחו.

אז אני נבחנתי במבחנים הבינלאומיים להסמכה בתחום של תכנון פיננסי (CFP) וחלק מהבחינה נשאלתי על משפט הגבול המרכזי שגורס כי מדגם מייצג הוא מדגם שגדול מ- 30 תצפיות כי אז המדגם עובר מהתפלגות בינומית להתפלגות נורמלית. אז אם מדגם של 30 תצפיות נחשב למדגם גדול, אז מדגם של 10,000 תצפיות לא מספיק?

מה אומר משפט הגבול המרכזי (Central Limit Theorem)? משפט הגבול המרכזי אומר שאם מחברים הרבה משתנים שבלתי תלויים זה בזה, ולא משנה כרגע איך כל אחד מהם מפולג (בינומית, פואסונית, גיאומטרית וכו') – הרי שהסכום שלהם מפולג נורמלית. במילים אחרות, סכימה של הרבה מאוד משתנים מקריים בלתי תלויים תביא לקבלת התפלגות נורמלית, כאשר לא משנה ההתפלגות של כל משתנה בודד, אלא מה שחשוב זה שמדובר במשתנים מקריים ושיש הרבה מאוד משתנים כאלה. עכשיו הרבה מאוד זה ממש לא 30. אני לא אכנס לדברים מסובכים אבל אני אומר שעל מנת לקבל התפלגות נורמלית באמצעות שימוש בסימולציות מונטה קרלו מעריכי שווי ואקטוארים עושים שימוש ב- 10,000 הרצות (Trails).

מה זה סימולציית מונטה קרלו?

סימולציית מונטה קרלו היא שיטה לפתרון בעיות חישוביות באמצעות מספרים אקראיים .למרות המקריות שבמספרים האקראיים, השיטה מאפשרת להגיע לרמת דיוק נדרשת על ידי שימוש בחוק המספרים הגדולים. סימולציית מונטה קרלו בנויה על דגימה אקראית של גורמי סיכון מתוך התפלגות מתאימה, כאשר על סמך הדגימה נוצרים מסלולים (Paths) דמיוניים של שינויים בגורמי סיכון לאורך זמן. רמת הדיוק של סימולציית מונטה קרלו נמדדת על ידי פרמטר שנקרא Standard error of sample mean אשר מהווה אינדיקציה לאיכות התוצאה.

תסביר לי מדוע צריך עשרות אלפי הרצות?

מכיוון שההתכנסות של סימולציית מונטה קרלו ל"ערך האמיתי" היא בסדר גודל של $1\sqrt{N}$, כאשר $\sqrt{N}$ הוא מספר ההרצות שנקבעו. פירושו של דבר, שכדי להגדיל את רמת הדיוק של המודל פי 10 אני נדרש לבצע עוד 100 סימולציות נוספות. לכן אקטוארים, מעריכי שווי ומנהלי סיכונים משתמשים ב- 10,000 הרצות.

אני ראיתי מעריכי שווי שמשתמשים בפחות. למשל בכמה הערכות שווי של אופציות אני ראיתי שמעריכי שווי הריצו סימולציית מונטה קרלו עם 5,000 הרצות בלבד.

ובכל זאת אתה רואה שלמרות שמעריכי השווי הללו מכירים את משפט הגבול המרכזי הם לא חשבו ש- 3,000 הרצות זה מספיק, קל וחומר ש- 300 הרצות זה מספיק וקל וחומר בנו של קל וחומר ש- 30 הרצות זה מספיק.

אז אתה אומר שאם יש לי 100 שורות בלבד של לקוחות ואני מחלק אותם ל- 70 שורות לסט אימון ו- 30 שורות לסט בדיקה, זה לא מספיק בעיניך?

התחושה שלי היא שזה פשוט לא יהיה מספיק נתונים.

מבחינה פרוצדורלית מול הלקוח, איך זה עובד פרויקט שכזה?

אז הצורה שבה אני כמדען נתונים עובד היא שאני עושה פגישת היכרות אחת ללא תשלום עם הלקוח. אני מגיע לפגוש אותו ושואל אותו המון שאלות על הנתונים הזמינים וכאלה דברים, ואם אני מוצא שיש טעם להמשיך, אז אני קובע משהו שנקרא בעגה של מדעני הנתונים "ימי הכנה". ימי הכנה אלו ימי הסתכלות על הנתונים, שזה כן בתשלום.

כמה ימים צריך כדי להסתכל על הנתונים?

את זה אני קובע בדרך כלל לפי המפגש הראשון, כמה זמן צריך להסתכל על הנתונים. אם זה באמת לא הרבה נתונים אז כנראה שמדובר על חצי יום עבודה. אבל החצי יום עבודה הזה הוא בתשלום.

על מה אתה מסתכל בנתונים?

אני רוצה לראות את הנתונים, האם מדובר בטבלה אחת או בהרבה טבלאות או דברים כאלה. אני אוסף אינפורמציה על מנת להבין איך הנתונים מגיעים, טבלה אחת, הרבה טבלאות, כמה שדות, כמה עמודות, כמה שורות, מה יש? מה אין? האם הוא יכול לתת לי דגימה של 5 שורות. הדגימה צריכה לכלול הן נתונים על לקוחות שנטשו והן נתונים על לקוחות שלא נטשו.

מה קורה אם אין לו תוויות?

ראשית נסביר לקורא מה זה תוויות (Labels). כאשר מדברים על נתונים בהקשר של מדע נתונים אז יש לנו שלושה מושגים. הראשון הוא מאפיין (Feature) שהוא משתנה המשמש באלגוריתם של למידת מכונה כמשתנה מסביר. המושג השני הוא יעד (Target) שהוא המשתנה שעליו אנחנו רוצים לעשות תחזיות. והמושג השלישי הוא תווית (Label) שהיא למעשה הערך של היעד. תוויות אלו נתונים מספריים על היעד. למשל, אם אני בונה מודל לחיזוי נטישת לקוחות, אז התוויות שלי הן התוצאה הסופית של כל לקוח, האם הוא נטש או לא נטש. חשוב שהלקוח שלי יבין שהוא זה שחייב לספק לי את הנתונים הללו. אם אין לי נתונים על לקוחות שנטשו אותו ואם אין לי נתונים על לקוחות שנשארו אז אין לי סט נתונים נורמלי. הרי אני לא אשב ואנחש לו ככה סתם מי יעזוב ומי לא. אני כמדען נתונים לא צריך לחפור לו בתוך המערכת ולהוציא את זה. הלקוח צריך לעשות את זה עבורי.

מה לאחר מכן?

לאחר מכן אני אומר ללקוח תודה אני צריך לחשוב ולאחר יום יומיים אני שולח לו הערכת זמנים ועלויות לפרוייקט. פשוט בלי להכיר את הנתונים ברמה הראשונית זה לא נראה לי אחראי לתת הערכה.

ככה בדיוק גם אני פועל כמעריך שווי וכשמאי רכוש. אתה שוב מחזיר אותי למבחנים הבינלאומיים להסמכה בתחום של תכנון פיננסי (CFP). אני זוכר שנבחנתי על אקונומטריקה ושם יש שני סוגים של נתונים: נתוני זמן ונתוני חתך-רוחב. האם נכון יהיה לומר שבמדע נתונים להבדיל מאקונומטריקה עובדים על וריאציה של נתוני זמן ולא על נתוני חתך רוחב? כלומר נגיד שאני עסק קטן שרוצה שאני שתבנה לו מודל לחיזוי לטישת לקוחות ויש לי בקושי נתונים על 100 לקוחות, ונניח שקיימת זמינות נתונים של עסק דומה עם נתונים על 50,000 לקוחות, האם תוכל להשתמש בנתונים של העסק הדומה על מנת לבנות מודל חיזוי לעסק הקטן שלי?

אני מבין למה אתה מתכוון ואני אנסה להסביר זאת גם לקורא. אתה בא ואומר שכחוקרים באקדמיה רוצים לבנות מודל ניבוי הם עושים שימוש בנתוני חתך-רוחב (Cross-Section), כלומר, אם הם רוצים למשל למצוא מודל ניבוי למכפיל ההון העצמי למשלל אז הם יריצו רגרסיות על נתונים של הרבה מאוד חברות כדי לבנות מודל ניבוי שכזה, בעוד שמדעני נתונים יבנו מודל חיזוי ספציפי לחברה המוערכת ולכן הם ישתמשו בנתונים של החברה המוערכת בלבד (Idiosyncratic Data). אני חושב שאתה צודק, כי מדעני הנתונים אכן מנסים כיום "לתפור" משהו שהוא מותאם בדיוק למידותיו (Tailor-Made) של הלקוח הספציפי כי פשוט אין כיום בעולם איזשהו מודל שבאופן גורף (Across the Board) עובד ומתאים לכולם. אין דבר כזה. לגבי מה שהצעת, אני יכול לקחת נתונים של לקוח אחר אבל אז אני צריך לבצע עליהם כל מיני התאמות לגודל, כיוונונים, קליברציות, כיולים, נרמולים וכדומה על מנת להתאים אותם ללקוח הספציפי.

אז אתה אומר למעשה שכן ניתן לעשות שימוש בנתונים של לקוח אחר עבור לקוח ספציפי?

בשביל זה יש מאמרים. ברגע שיש פרוייקט מסוים של מדע נתונים אז אני כמדען נתונים יושב וקורא קצת באינטרנט, רואה מה קיים? ממה אני יכול ללמוד? בדרך כלל המצב הוא שאני פשוט צריך לעשות בעצמי הנדסת מאפיינים (Feature Engineering) אבל לרוב רני שואב רעיונות מדברים שאני קורא. אתה חייב להבין בדרך כלל הסטים של הנתונים הם מאוד לוקאליים כאלה, הם מאוד מאופייני פרוייקט ולכן אני בונה לעצמי את הנתונים בצורה כזאת. למשל, אני בונה מאפיין לקוח, כל לקוח הופך להיות איזושהי רשומה עם 0 ו- 1. או אם זה בעיית רגרסיה של כמה זמן מראש הלקוח ינטוש אותך, אז אני אומר תוך כמה זמן הוא נטש, אני עושה איזשהו איזון של הנתונים, אני שואל את עצמי אני יוצר נתונים חדשים או אפילו חותך את הנתונים הקיימים.

מה זה האיזון הזה שציינת?

על מדעון הנתונים לייצר איזון כלשהו בין כמות הנתונים על הלקוחות שנטשו אותך לבין כמות הנתונים על הלקוחות שלא נטשו אותך. בקיצור, יש כאן הרבה שאלות שדורשות היכרות עם הדאטה ודיבור עליו וזה דורש את הזמן הזה. לא סתם אני אומר שככה מדעני נתונים עובדים.

מה הסדר גודל מבחינת עלות של פרוייקט בניית מודל לחיזוי נטישת לקוחות?

לא ניתן להעריך את שכר הטרחה. יש משימות שיכולות לקחת יום עבודה ויש משימות שיכולות לקחת הרבה זמן. זה מאוד תלוי איך הנתונים מגיעים. האם הם מסודרים ברמה שהם מוכנים להרצה או שצריך לעבוד עליהם הרבה. רב הפעמים הנתונים לא מסודרים ויש עוד נתונים שלא מופיעים במאגר וצריך לעשות כל מיני מניפולציות עליהם. שבוע ימים זה המינימום לפרוייקט שכזה והמקסימום הוא חודש עבודה.

אבל איך אתה עובד איתו שעתי או גלובלי?

אני בדרך כלל מתמחר שעתי, אבל הלקוחות לא אוהבים את זה. מה שאני בדרך כלל אומר ללקוח זה: "בואו תוציא לי הזמנת עבודה קטנה, 16 שעות, יומיים עבודה ואני אגיד לך מה הכיוון אם אתה רוצה, ואחרי זה נדבר". אבל תבין יכול להיות שבשלב הזה אני גם אומר לו שאין כלום. תבין יש כאן בעיה, זה לא שהוא מבקש הערכת שווי על בסיס דוחות כספיים, שזה אתה ישר יושב ועושה. כאן מדובר בתחום שיכול להיות שאי אפשר לבנות מודל שכזה או לחילופין שהמודל שאני אבנה ללקוח לא ימצא חן בעיניו כי הוא יתן לו 30% שגיאה ואז הלקוח יגיד לי שזה לא עוזר לא בחיים. מדובר בסוג של עבודות שבהן התשובה יכולה להיות בסוף שאין מודל. שאין תשובה. הלקוח צריך לדעת את זה ולהבין את זה. יבוא אליי עכשיו לקוח ויגיד לי אני רוצה מודל שיגריל את המספרים הנכונים בלוטו בהגרלה הבאה. נו, יופי, אז מה אם הוא רוצה. אמנם אנו עוסקים במודלי חיזוי אבל לא תמיד אפשר לחזות משהו. הרי אין לי כדור בדולח.

אבל ניתן למדוד את כושר הניבוי של המודל, ישנם כלים סטטיסטיים לכך.

אם בכלל. כי אם אני כמדען נתונים צריך לנבא האם לקוח מסוים ינטוש או לא ינטוש ונתתי ללקוח ניבוי ברמת ביטחון של 60%, אז זה לא ימצא חן בעיניך כי זה אומר שיש כאן 40% טעות.

עכשיו אני מבין למה פרוייקטים כאלה מתומחרים בשעות.

תראה, יש לקוחות שלא מסכימים למתווה שעתי ואז אני אומר להם שהפרוייקט ינוע בין שבועיים לחודש ושאני אהיה מספיק הוגן על מנת לתת להם האם זה הולך לכיוון של השבועיים או לכיוון של החודש. אם אני רואה שזה משהו שאין לו פתרון או שזה משהו שנפתר בקלות אז זה ייגמר בתוך שבועיים ואז אני אקח מה שצריך. אני לא מחפש לגנוב לקוחות. המינימום זה שבועיים והמקסימום זה חודש.

מה הלקוח מקבל בסוף אם אין פתרון?

כך או כך הלקוח מקבל בסוף דוח מלא של האנליזה של כל מה שעשיתי, עם כל החיתוכים, כל הדברים, כל הניסיונות. דוח מלא של כל מה שנעשה.