

מדען נתונים, על ניקוי נתונים שמעת כבר?

ניקוי נתונים הינו היבט חשוב ביותר, שלא לומר מרתק, של למידת מכונה. מסקרים שנעשו בארה"ב על הנושא עולה כי מדעני נתונים מקדישים לא פחות מ- 80% מזמנם לניקוי נתונים.

ברגיל, מאגרי מידע גדולים טומנים בחובם בעיות הטעונות תיקון. ניקוי נתונים טוב הוא זה שעושה את כל ההבדל בין למידת מכונה מוצלחת לכזו שאינה מוצלחת. הביטוי "זבל נכנס, זבל יוצא" (out-garbage-in, garbage) רלוונטי וישים לגבי למידת מכונה ממש באותה מידה שהוא תקף ונכון לגבי ניתוחים אחרים.

בשלב זה, נציין כי קיימים שני סוגים של נתונים: נומריים (Numerical) וקטגוריים (Categorical).

נתונים נומריים מכילים מספרים, בעוד שנתונים קטגוריים הינם נתונים שיכולים 'ליפול' לתוך מספר קטגוריות שונות. לדוגמא, נתונים לניבוי מחירי דירות יכולים לקטלג דירה מסוימת כמשופצת או כלא משופצת, עם או ללא מעלית, עם מזגן או בלי מזגן וכיוצא באלה קטגוריות (שותפה מעשנת, שומרת שבת, נקיה, אוהבת חיות ומסיבות וכו'). למעשה ניתן להמיר נתונים קטגוריים למספרים לצורך הניתוח.

הסוגיה הראשונה הקשורה לניקוי נתונים הינה "תיעוד או רישום לא עקבי" (Inconsistent Recording) של נתונים. לאמור- נתונים נומריים וקטגוריים כאחד עלולים להיות חשופים לתיעוד לא עקבי. לדוגמא, נתונים נומריים עבור שטח של דירה מסוימת עשויים להיות רשומים במאגר נתונים כלשהו כ- 100, מאה, 1.0 ממ"ר, 100 מ"ר, +100, וכך הלאה. לפיכך, חשוב ביותר ראשית לבחון את הנתונים על מנת לקבוע איזו וריאציות יכולים לקבל הנתונים ואז להחליט על הגישה הטובה ביותר לניקוי. נתונים קטגוריים עשויים לציין דירה מסוימת כ-: "משופצת", "אחרי שיפוץ", "כמו חדשה" וכו'. הגישה הפשוטה ביותר היא להכין רשימה של כל החלופות הקיימות במאגר



עבור מאפיין מסוים (משופצת או לאו במקרה דנן שלפנינו) ולאחר מכן לאחד אותם בצורה הולמת.

הסוגיה השנייה הקשורה לניקוי נתונים מכונה "תצפיות לא רצויות" (Unwanted Observations). לדוגמא, אם אתה מפתח מודל לניבוי מחירי דירות באזור מסוים, אז חלק מהנתונים שלך עשויים להתייחס בכלל למחירי דירות או מחירי בתים שאינם באזור הנבדק. על כן חשוב מאוד למצוא דרך לזהות את הנתונים הללו ולהסירם טרם התחלת הניתוח.

הסוגיה השלישית הקשורה לניקוי נתונים נקראת "תצפיות כפולות" (Duplicative Observations). לדוגמא, כאשר נתונים ממספר מקורות מידע שונים "מוכנסים" לתוך מאגר נתונים אחד, הרי שאך טבעי שתהיינה תצפיות כפולות אשר עלולות כידוע להביא להטיית תוצאות הניתוח. לפיכך, חשוב לעשות שימוש באלגוריתמים של חיפוש על מנת למצוא ולהסיר כפילויות ככל שרק שניתן.

הסוגיה הרביעית הנוגעת לניקוי נתונים נקראת "חריגים" (Outliers). במקרה של נתונים נומריים ניתן לזהות חריגים על ידי העלאת הנתונים על תרשים (תרשים פיזור או היסטוגרמה) או לחילופין באמצעות חיפוש נתונים אשר מרוחקים שש סטיות תקן (Six Sigmas) מהתוחלת. לעיתים די ברור שהחריגים הינם תוצאה של טעות הקלדה. לדוגמא, אם שטחה של דירת 3 חדרים מופיע במאגר כ- 1,000 מ"ר אז די ברור שמדובר בטעות הקלדה. עם זאת, כפי שלימד אותי מורי הרבי המלומד ה"ה ד"ר שילה ליפשיץ ז"ל (שהייתה לי הזכות להיות עוזר מחקר שלו במשך 4 שנים) "חריגים מסירים רק אם קיימת סיבה מספיק טובה לעשות זאת". נתונים נומריים שגדולים באופן יוצא דופן או שקטנים בצורה לא רגילה, אם הם אכן נכונים, הרי שהם מכילים מידע שימושי. השפעתם של חריגים על התוצאות של למידת מכונה תלויה במודל שבו נעשה



שימוש. כך למשל לחריגים ישנה השפעה רבה על מודלים מסוג רגרסיה, בעוד שעל מודלים מסוג עצי החלטה (Decision Trees) כמעט ואין להם כל השפעה.

הסוגיה החמישית הנוגעת לניקוי נתונים מכונה "נתונים חסרים" (Missing Data). כמעט בכל מאגר נתונים גדול חסרים ערכים. הגישה הפשוטה לטיפול בסוגיה זו היא פשוט להסיר את הנתונים שלגביהם חסרים ערכים עבור מאפיין אחד או יותר. אבל זה ממש לא רצוי הואיל ופעולה שכזו מקטינה אפקטיבית את גודל המדגם ועשויה ליצור הטיות. במקרה של נתונים קטגוריים, הפתרון הפשוט ביותר הוא ליצור קטגוריה חדשה שנקראת "חסרים" (Missing).

במקרה של נתונים נומריים, אחת הגישות היא פשוט להחליף את הנתונים החסרים בתוחלת או בחציון של ערכי הנתונים שאינם חסרים. לדוגמא, אם שטחה של דירה מסוימת חסר ומצאנו שחציון שטחי הדירות ישנם נתונים זמינים נאמד בכ- 115 מ"ר, הרי שאנו יכולים לעשות שימוש בערך שמצאנו עבור כל יתר הדירות שחסרים לנו נתונים אודות גודל שטחן.

גישות קצת יותר מתוחכמות כוללות הרצת רגרסיות על היעד (Target), מחירי הדירות במקרה דנן (שלפנינו) ביחס לערכים שאינם חסרים ולהשתמש בתוצאות הרגרסיה עבור הערכים שחסרים. לעיתים מקובל להניח שנתונים חסרים באופן מקרי (פשוט כי ככה זה) ולעיתים עצם העובדה שחסרים נתונים היא כשלעצמה אינפורמטיבית. במקרה האחרון, מקובל לייצר משתנה אינדיקטור חדש שמקבל את הערך 0 אם קיימים נתונים או את הערך 1 אם חסרים נתונים.

פרטים אודות כותב המאמר: האקטואר רועי פולניצר, FRM

רועי בעל תואר שני במימון (התמחות בניהול סיכונים ואקטואריה) ותואר ראשון בכלכלה (התמחות במימון), שניהם מאוניברסיטת בן-גוריון בנגב, בעל דיפלומה בניהול סיכונים פיננסיים (FRM®) מאוניברסיטת אריאל בשומרון ולמד בתוכנית ללימודי תעודה באקטואריה באוניברסיטת חיפה. כמו כן, רועי אקטואר מלא



(Fellow) בלשכת מעריכי השווי והאקטוארים הפיננסיים בישראל (F.I.L.A.V.F.A.), מוסמך כמעריך שווי מימון תאגידי (CFV) מטעם לשכת מעריכי השווי והאקטוארים הפיננסיים בישראל (IAVFA), מוסמך כמנהל סיכונים פיננסיים (FRM) מטעם האיגוד העולמי למומחי סיכונים (GARP) ומוסמך כמומחה לניהול סיכונים (CRM) מטעם האיגוד הישראלי למנהלי סיכונים (IARM).

לרועי ניסיון של מעל ל- 15 שנה בביצוע ניתוחים כמותיים במכשירים פיננסיים, בהערכת שווי תאגידים ונכסים בלתי מוחשיים, באמידה וכימות סיכונים כמו תמותה, אריכות ימים, תחלואה, ביטולים והחלמה מנכות, ובמידול ומדידת סיכוני שוק, אשראי, תפעוליים, מודל, מזילות והשקעות לצורכי יישום הוראות רגולטוריות ותקינה חשבונאית, פיתוח, יישום ותיקוף מודלים בתחומים של הערכות שווי, ניהול סיכונים, אקטואריה והנדסה פיננסית, קביעת תעריפי ביטוח חיים, הערכת פרמיות סיכון והערכת עתודות ביטוח, קביעת עלות תנאי פנסיות (צוברות ותקציביות) והכנת מאזנים אקטואריים לקרנות פנסיה, ניתוח וחזוי מצבים פיננסיים מורכבים וכן העברת סמינרי הדרכה והשתלמויות בתחומי התמחות: מימון, אקטואריה, הערכות שווי, בנקאות, ניהול סיכונים, אופציות והנדסה פיננסית.



ניסיונו של רועי בתחום ה-Data Analysis, כולל: עבודה עם מאגרי מידע גדולים Big Data תוך שימוש ב-Statistical Learning (כגון: סטטיסטיקה תיאורית, הסתברות, הסקה סטטיסטית, סטטיסטיקה א-פרמטרית, חלוקת נתונים, נרמול נתונים, Fitting ו- Bayes Theorem) ובאלגוריתמים מסוג Unsupervised Learning (כגון: Hierarchical Clustering, k-means Clustering, Density-based Clustering, Distribution-based Clustering ו- Principle Components Analysis) למציאת דפוסים וזיהוי מגמות ואנומליות בעולמות ניהול הסיכונים, ההשקעות, האקטואריה, הביטוח והפנסיה, פיתוח תשתית לצורך ניתוח נתונים, שילוב והטמעת כלים לצורך גישה ושליפה עצמאית של נתונים ממאגרי מידע, פיתוח דוחות, ממשקים ומסכים באמצעות כלי ויזואליזציה.

ניסיונו של רועי בתחום ה-Data Science, כולל: עבודה עם מסדי נתונים גדולים Big Data תוך שימוש באלגוריתמים מסוג Supervised Learning (כגון: Linear Regression, Ridge Regression, Lasso Regression, Elastic Net Regression, Logistic Regression, Maximum Likelihood Estimation, k-Nearest Neighbors, Decision Tree, Random Forest, Ensemble, Bagging, Boosting, Naïve Bayes Classifier, Linear Separation, Support Vector Machine, Non-Linear Separation, SVM Regression, Artificial Neural Network, Convolutional Neural Network ו- Recurrent Neural Network) לניבוי וסיווג בעולמות ניהול הסיכונים, ההשקעות, האקטואריה, הביטוח והפנסיה ובמודלים מסוג Reinforcement Learning (כגון: Q-learning, Monte Carlo, Simulation, Temporal Difference Learning ו- n-Step Bootstrapping) לקבלת החלטות מרובות שלבים בעולמות ניהול הסיכונים, ההשקעות, האקטואריה, הביטוח והפנסיה, זיהוי אתגרים עסקיים שבהם DATA יכול להוות גורם מכריע בשיפור קבלת החלטות, איתור ואיסוף מקורות מידע, הגדרה ואיפיון של שימושי המידע, בניית מסד המידע, אפיון והגדרת הצגת המידע



ותוצריו, פיתוח כלים, מודלים, תהליכים ומערכות בתחום האנליזה, תוך שימוש בכלי אנליזה מתקדמים (EXCEL, VBA ו-R).