כמה פעוטות יהודים זכרים בני שנתיים אתם מכירים שמתו בחודש יולי 2022? אני יודע כמה היו יכולים למות!

האקטואר רועי פולניצר מסביר כיצד בנה אלגוריתם לניחוש מספרי הגרלת הלוטו וכיצד הצליח 5 הגרלות ברציפות לנחש 3 מספרים מתוך 6 המספרים הנדרשים, באמצעות האלגוריתם שבנה.

בחודש מאי השנה שאל אותי חבר האם אני כ- Data Scientist (כלומר אלגוריתמאי שלמעלה מ- 17 שנים בונה מודלים מתמטיים וסטטיסטיים לזיהוי דפוסי התנהגות חוזרים) מסוגל לבנות אלגוריתם לחיזוי מספרי הלוטו. מכאן מתחילה הכתבה שלי.

תחילה, נזכרתי שכאשר למדתי אקטוארריית ביטוח חיים ופנסיה בתוכנית ללימודי תעודה בחוג לסטטיסטיקה באוניברסיטת חיפה, עשינו קצת חישובים שקשורים לזכייה בלוטו. בשבוע החמישי ללימודים, בשנה הראשונה, למדנו שהגרלת הלוטו היא הגרלה מקרית ושהיא מפולגת היפר-גיאומטרית.

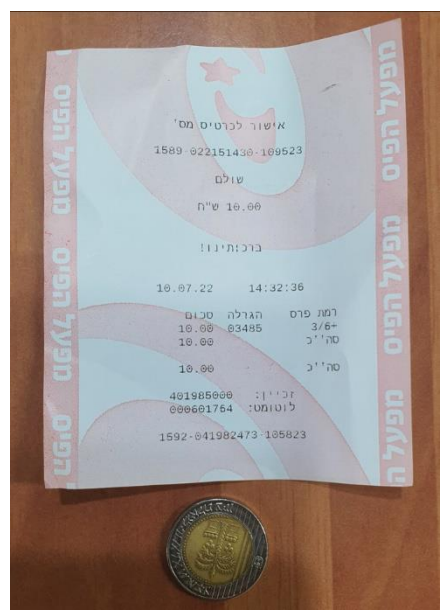
באותו שיעור על התפלגות היפר-גיאומטרית, חישבנו את הסיכוי לזכות בפרס הראשון (לנחש נכון 6 מספרים מתוך 37 מספרים ללא החזרה וגם לנחש מכון מספר חזק מתוך 7 מספרים – סיכוי של 0.00000614%), בפרס השני (לנחש נכון 6 מספרים מתוך 37 מספרים ללא החזרה – סיכוי של 0.00004301%) ובפרס השלישי (לנחש נכון 5 מספרים מתוך 37 מספרים ללא החזרה וגם לנחש מכון מספר חזק מתוך 7 מספרים – סיכוי של 0.00003277%).

למעשה, אותו חבר ביקש ממני לפתח אלגוריתם לזכייה בפרס הראשון בלוטו, שזה סיכוי של 1 ל- 16,273,488. אבל הסיכוי הזה נכון רק בתנאי שמדובר בהגרלה מקרית ושאין שום טריקים ושטיקים בהגרלת הלוטו. נשאלת השאלה, האם הגרלות הלוטו הן אכן הגרלות מקריות?

כאלגוריתמיקאי אני יודע שאם מדובר בהגרלות מקריות, אז אין "מודל חבוי" בתוך נתוני הגרלות הלוטו. מה זה "מודל חבוי"? הכוונה היא לכלל/קשר כלשהו (לינארי או אחר) שמסתתר בתוך הנתונים ושאם נעלה עליו נוכל לבצע פרידקציות יעילות ולחזות התנהגויות עתידיות ברמת דיוק גבוהה.

עד היום כל הפרוייקטים של מדע נתונים שעשיתי אכן מצאתי בנתונים "מודל חבוי". חשוב להבין, שאם הנתונים הם אכן מקריים, אז אין מודל חבוי בנתונים ואם אין מודל חבוי בנתונים הרי שלא אצליח לבנות מודל חיזוי טוב.

ואז נזכרתי שכאשר שלמדתי אקטוארריית ביטוח כללי בתוכנית ללימודי תעודה במכללת ג'ון ברייס, המרצה שלי סיפר לנו בשיעור על סדרות עתיות (סדרות עתיות פירושן שאין מקריות בנתונים) על כך שבאתר הלוטו יש את כל התוצאות שיצאו אי פעם בהגרלות ושהרבה אנשים ניסו את מזלם באמצעות הנתונים הללו. כלומר, הרבה אנשים אמרו לעצמם "בוא ננסה לעשות מזה מודל, אולי נצליח לנחש מה יהיו המספרים של הלוטו. ואם לא יצא מודל שמנחש את כל 4 המספרים, אז מספיק לנחש רק 3 מספרים מתוך ה- 6 וזה כלשעצמו כבר מצמצם את כמות האפשרויות על מנת לזכות בפרס הראשון".



אותו מרצה קבע אקסיומטית שאי אפשר לבנות מודל לניחוש מספרי הלוטו כי אין מודל חבוי בנתוני ההגרלות הואיל ומדובר בנתונים אקראיים לחלוטין. אז החלטתי שמעניין אותי לבדוק בעצמי את קביעתו של המרצה. אולי לא מדובר בנתונים אקראיים. אולי פשוט עד היום אף אחד לא היה מסוגל למצוא את המודל החבוי בנתונים, כי זה פשוט היה מסובך מדי למצוא אותו.

אז ניגשתי לארכיון תוצאות הגרלות הלוטו ומשכתי משם את כל ההגרלות השבועיות מה- 3 בספטמבר 1968 ואילך. הנתונים מסודרים בקובץ csv (ורסיה של קובץ אקסל) בצורה טבלאית כאשר בכל הגרלה, המספרים שעלו בגורל רשומים בסדר עולה מהנמוך ביותר לגבוה ביותר. כך למשל, בעוד שבהגרלת הלוטו שנערכה ב- 19.11.2022 (מצולמת וניתנת לצפייה באתר הלוטו) 6 המספרים שעלו בגורל לפי סדר הגרלתם היו 30, 13, 35, 10, 29 ו- 16, הרי שהמספרים המופיעים בקובץ ה- csv מסודרים תחת 6 קטגוריות: כדור מס' 1, כדור מס' 2,... עד כדור מס' 6 והסדר שבו הם מוצגים הוא 10, 13, 16, 29, 30 ו- 35.

מהנתונים עולה שכדור מס' 1 (כלומר, הכדור הראשון בהגרלה) הוא תמיד המספר הנמוך ביותר בכל ההגרלות בעוד שכדור 6 כלומר, הכדור השישי בהגרלה) הוא תמיד המספר הגבוה ביותר בכל ההגרלות. על פניו, נראה שיש דפוס חוזר בנתונים שרק מחכה שאבוא ואמצא אותו.

לאחר שניתחתי את הנתונים נתגלו בפניי כמה אנומליות שלאחר בירור עם חברים חובבי לוטו מסתבר שכללי ההגרלה בארץ השתנו מאז שנת 68 ועד היום. למשל, כיום 6 המספרים שיש לנחש הם מספרים טבעיים (מספרים שלמים וחיוביים) שיכולים לקבל ערכים בין 1 ל- 37 (כולל), אבל עד ל- 9.3.2004 (כולל) 6 המספרים יכלו לקבל ערכים בין 1 ל- 49 (כולל).

לכן מחקתי ממסד הנתונים שלי את כל ההגרלות שנערכו מה- 3.9.1968 ועד ה- 9.3.2004 (כולל). בנוסף, כיום המספר החזק הוא מספר טבעי (מספר שלם וחיובי) שיכול לקבל ערכים בין 1 ל- 7 (כולל), אבל עד ל- 1.3.2011 (כולל) המספר החזק יכל לקבל ערכים בין 1 ל- 10 (כולל). ולכן נאלצתי למחוק גם את כל ההגרלות מיום ה- 12.3.2004 ועד ל- 1.3.2011 (כולל).

מאחר שאני מעוניין לאמן מודל של רשתות נוירונים והואיל וידוע שרמת הדיוק של רשתות נוירונים עולה כאשר מדובר בנתונים מנורמלים בעלי תוחלת של 0 ושונות של 1, על לכן למרות שמספרי הכדורים המקוריים נעים בין 1 ל- 37, ביצעתי נירמול למספרים כך שיהפכו ממספרים טבעיים למספרים ממשיים (מספרים לא שלמים שיכולים להיות חיוביים או שליליים) בגדול בין 3- ל- 3+.

משעה ש"כיוונתי" את מסד הנתונים המקורי שלי להיות בעל התפלגות נורמלית סטנדרטית, בניתי את מסד הנתונים החדש שלי כך שכל 7 שורות "יודבקו" לשורה אחת (למעשה מסד הנתונים שלי הצטמצם פי 7).

במילים אחרות, בניתי את מסד הנתונים שלי כך שמשתני הכניסה (Features) שלי יהיו מספרי 7 ההגרלות שקדמו להגרלה העוקבת להן על פי סדר הגרלתן. כך למשל, השורה הראשונה של מסד הנתונים החדש שלי מורכבת מ- 49 משתני כניסה (7 הגרלות כפול 7 מספרים בכל הגרלה: 6 מספרים רגילים ו- 1 מספר חזק), הנגזרים מנתוני 7 ההגרלות הראשונות במסד הנתונים החדש שלי: 5.3.2011, 8.3.2011, 12.3.2011, 15.3.2011, 19.3.2011, 22.3.2011 ו- 26.3.2011 לפי הסדר.

בנוסף, בניתי את מסד הנתונים שלי כך שמשתני היציאה (Target) שלי יהיו מספרי ההגרלות העוקבת ל- 7 ההגרלות שקדמו להן. כך למשל, השורה הראשונה של מסד הנתונים החדש שלי מורכבת מ- 7 משתני יציאה (הגרלה אחת כפול 7 מספרים בכל הגרלה: 6 מספרים רגילים ו- 1 מספר חזק), הנגזרים מנתוני ההגרלת השמינית במסד הנתונים החדש שלי: 29.3.2011.

כדי לתקף את המודל שלי על התוצאות בפועל – חילקתי את נתוני המסד החדש לשתי קבוצות. הקבוצה הראשונה לה קראתי קבוצת ה"אימון" (train set) שימשה אותי לאימון המודל (כל ההגרלות מבלד 8 ההגרלות האחרונות/העדכניות ביותר). הקבוצה השנייה לה קראתי קבוצת ה"ביקורת" (test set) שימשה אותי לתיקוף המודל (8 ההגרלות האחרונות / העדכניות ביותר בלבד).

מסתבר שהמודל שלי יודע לנחש נכון **לפחות** 3 מתוך 6 המספרים בלוטו (בדגש על המילה לפחות כי לעיתים הוא מנחש גם 4 מספרים ו- 5 מספרים נכון, וזה מבלי לקחת בחשבון את המספר החזק). רק כדי לסבר את האוזן, הסיכוי לנחש נכון 3 מספרים מתוך 6 מספרים בלוטו, כאשר המספרים נעים בין 1 ל- 37 כולל ללא החזרה וללא מספר חזק הוא בערך 0.013% (ליתר דיוק סיכוי של 1 ל- 7,770).

בחודש יולי האחרון (2022) זכיתי ב- 5 הגרלות רצופות, כאשר בכל אחת מהן זכיתי בפרס ה- 16 (כלומר, ניחשתי נכון 3 מספרים מתוך 6 – סיכוי של בערך 0.013%, זוכרים?). מה היה גובה הפרס? פעם אחת 10 ש"ח וביתר הפעמים 15 ש"ח בבכל פעם (תעשו תחשבון לבד).

למי שרוצה לקבל המחשה ממשית להישג של המודל שבניתי אציין שהסיכוי לזכות בפרס ה- 16 בהגרלת הלוטו שקול אפקטיבית מכל הבחינות הסטטיסטיות המהותיות להסתברות שפעוט יהודי זכר בן שנתיים (2) לא יגיע בחיים ליום ההולדת 3 שלו – סיכוי של 0.013% על פי לוחות תמותה שלמים של ישראל 2016–2020 שפורסמו ב- 9 במאי 2022.

לוח 5. לוח תמותה שלם של ישראל:
יהודים – זכרים
2016-2020

נשארים בחיים בגיל x Survivors at age x l_x	הסתברות למות Probability of death: רווח סמך Confidence interval – גבול עליון Upper boundary	גבול תחתון Lower boundary	סטיית תקן Standard deviation	q_x	גיל Age
100,000	0.00247	0.00215	0.00008	**0.00231**	0
99,769	0.00022	0.00014	0.00002	**0.00018**	1
99,751	0.00018	0.00008	0.00002	**0.00013**	2
99,738	0.00013	0.00007	0.00002	**0.00010**	3
99,728	0.00011	0.00005	0.00002	**0.00008**	4

כזכור, המודל שלי הצליח 5 פעמים ברציפות להביא אותי לתוצאה הזו במהלך חודש אחד. ואני שואל, את מי שאומר שהגרלות הלוטו הן מקריות, **כמה פעוטות יהודים זכרים בני שנתיים אתם מכירים שמתו (חלילה וחס) בחודש יולי האחרון?**

ולמי שחושב שהסיפור שלי הזוי, אז מצ"ב קישור למאמר מדעי שלי באנגלית שבמסגרתו פיתחתי אלגוריתם מתמטי וכתבתי אותו בקוד פייתון שכתבתי בסוף מאי 2022, לפני שהתחלתי "לעשות כסף" מהמודל:

Polanitzer, R (2002). *How to Guess Accurately 3 Lottery Numbers Out of 6 using LSTM Model*, Medium.

הכותב הוא הבעלים ומדען הנתונים האחראי של "פרדיקציות יועצים" משנת 2010 ועד היום. משנת 2005 מיישם שיטות אנליזה לתחקור מיד ובונה מודלים מתמטיים וסטטיסטיים לזיהוי דפוסי התנהגות חוזרים ולחיזוי התנהגויות עתידיות לצורך הפקת תובנות עסקיות.

הנושאים בהם מר פולניצר עוסק: ביצוע מחקרי מידע מעמיקים לטובת הפקת תובנות עסקיות לארגונים, ניקוי, טיוב וסידור מידע המשמש למחקרים שונים, הפעלת אלגוריתמים שונים של מידול, כריית מידע ו- Machine Learning על מידע, בניית תהליכי הכנת מידע ואופטימיזציה של אלגוריתמים שונים וכו'.

חבר מלא באיגוד הישראלי למדעני נתונים מקצועיים וזכה במקום הראשון בתחרות הדאטה סיינטיסטים של מכללת ג'ון בריס ה- Maximum F1-Score.