בס"ד

**סיכום הפגישה עם אלה – 11.6.25**

יש להכניס בשלב יצירת הדאטהסט – התנהגות שונה קבוצתית המותאמת לסוג התפקיד. לדוגמה: המזכירות מדפיסות יותר דפים בהשוואה לבעלי תפקידים אחרים. רק כך הדאטה יהיה קרוב למציאות.

ההתייחסות לקבוצה (לתפקיד) חייבת להיות לא רק בשלב יצירת הדאטה אלא גם בשלב ההתייחסות לדאטה. לדוגמה, אם ננסה למצוא אנומליה לפי כמות הדפסת הדפים – אנו בסיכון להחשיב את כל המזכירות כחשודות למרות שזו ההתנהגות הסטנדרטית לקבוצה זו.

איך נתמודד מול ההבדלים בין הקבוצות? עלינו לנרמל את ההתנהגות של המזכירה ביחס לשאר המזכירות. אנחנו מנרמלים את ההתנהגות של כלל הקבוצה, ובמזכירה הספציפית הזו את בונה את ה-time series ומזהה אנומליות. זה ביחס לתפקיד.

כשאת עושה נרמול להתנהגות של המזכירה - אם זוהי התנהגות תקינה לקבוצה את כבר לא תראי אנומליה. את זה עושים לפני שמכניסים את הנתונים למודל שעובר על ה- .time series במילים אחרות, את מוצאת את ההתנהגות הנורמלית כדי למצוא את החריגים שחרגו מהנורמה הזו.

הנרמול של הקבוצה הוא בהתאם לממוצע ולסטיית התקן שהיו בקבוצה באותו יום. למה הנרמול מתבצע לכל יום בפני עצמו? כי זה time series יש בעבודה תקופות עמוסות יותר ופחות, וגם ימים במהלך השבוע שיותר עמוסים מימים אחרים.

מבחינת קוד זה משהו קבוע: עושים לולאת תפקידים ובתוכה לולאת עובד ושומרת לכל תפקיד את הממוצע באותו יום ואז בודקת כל עובד ביחס לאותו ממוצע.

האם צריך להוסיף פיצ'רים בדאטה בייס שקשורים לנרמול?

לא, כשהמודל יקבל את הנתונים הוא כבר יקבל אותם מנורמלים. ולכן הוא יראה בתור אנומליה רק את הנתונים שהם מוזרים גם ביחס לקבוצה.

את הנרמול נעשה כחלק מה-preprocessing.

לזהות אנומליה של האדם ביחס להתנהגות שלו עצמו, כלומר לזהות שפתאום נהייתה לו התנהגות חריגה – את זה עושים עם מודלים כמו LSTM.

אין צורך לבצע clustering בשלב האימון, היות שאנו כבר יודעות מהן הקבוצות – מהם התפקידים הושנים. אולם כדאי לבצע clustering בשלב ה-EDA כדי להציג איך הדאטה מתפלג. נציג את ההתנהגות השונה של התפקידים השונים וכך נרוויח שתי מעלות: א. יהיה מובן לכולם מדוע היה צורך לבצע נרמול לפי תפקידים. ב. נראה כיצד יצרנו את הדאטה בהתאם לחיים האמיתיים. אולי גם נמצא קבוצות של תפקידים שמתנהלות זהה.

אפשר לתת ל-AI פרומפט שבו אומרים לו ליצור דפוסים על פי התפקידים השונים (ורק אחרי שנאשר את הדפוסים האלו הוא יג'נרט את הדאטה לפיהם), או לומר לו ברחל ביתך הקטנה: מזכירות עובדות בערך בשעות האלו וכו'.

אחרי שהוא מג'נרט את הדאטה – כדאי להוסיף רעש נורמלי כפי שהוספנו בעבר. אלה מוסיפהף כשמוסיפים רעש אפשר לעגל להכי קרוב.

\* בעת האימון לא כדאי לתת למודל את כל הפיצ'רים בהתחלה, על מנת שלא נטביע אותו עם יותר מידי מידע.

\* כדאי להתחיל לאמן עם דאטה לא מנורמל כדי לראות מה התוצאות בלי הנרמול, ואחר כך לנרמל ולהראות איך התוצאות השתפרו בעקבות כך. באופן דומה, כדאי להתחיל עם מודל פשוט כמו isolation forest, להתקדם אולי ל- Arimaולהמשיך עם מודלי deep learning, כדי להראות את השיפור העקבי בדיוק. היות שבפרויקט רוצים לראות שעשיתן פתרון פשוט לבעיה וכיצד שיפרתן אותו שוב ושוב.

\*שאלה: היות שיש דפוסי עבודה עונתיים שקשורים לקבוצה (לדוגמה תקופת לחץ שבה יש לכל המזכירות יותר הדפסות), האם יש צורך לעשות time series על כל תפקיד בפני עצמו, או שהנרמול ביחס לממוצע ולסטיית תקן באותו יום – כבר מייתר את הצורך ב- time series?