פרויקט חלק ב'

משימה 2 - Learning Task

Pre-Processing לדאטה לצורך למידה

1. **הורדת משתנים שאינם תורמים ללמידה**

א. במסגרת החלק הראשון ראינו שישנן מספר עמודות שניתן לוותר עליהן לצורך הלמידה של העמודה gt:

1) $Model$ – העמודה מכילה ערך אחד בלבד, לכן אינה תורמת לחיזוי.

2) $Creation\_Time$ – ראינו כי אין הבדל משמעותי בין הזמנים המופיעים בעמודה זו לבין הזמנים בעמודה $Arrival\_Time$ (ראה מסקנה 2 בחלק הראשון של הפרויקט להסברים נוספים).

3) $Index$ – עמודה זו אינה אלא מספור הרשומות.

2. **משתנים קטגוריאליים ← משתני dummy**

א. כדי שנוכל להתייחס למשתנים שאינם רציפים כפיצ׳רים בעלי משמעות במשימת הלמידה, ביצענו טרנספורמציה לכל משתנה קטגוריאלי.

ב. לכל משתנה קטגוריאלי יצרנו עמודות, כמספר הערכים האפשריים שמשתנה זה מקבל.

ג. לכל רשומה במסד, קיבלנו עבור כל משתנה מספר עמודות המכילות 0 ועמודה אחת המכילה 1, המסמלת שהערך של המשתנה הוא הערך המציין את העמודה.

ד. המשתנים הקטגוריאליים עליהם ביצענו את הפעולות הנ״ל - $Device, User$.

ה. בנוסף, נציין כי את העמודה gt שהכילה ערכים מילוליים קודדנו למספרים שלמים (לא $dummies$, אלא קידוד בלבד). עמודה זו היא ה-$label$, וקידוד זה נעשה לצורך נוחות.

3. **נרמול משתנים נומריים לפי נוסחת $min - max$**

א. למעט פיצ׳רים שראינו שניתן להורידם טרם הלמידה (בהתאם למסקנות החלק הראשון), לא הגענו למסקנות חותכות על אודות פיצ׳ר/ים מסויימ/ים בעל משקל שונה מהאחרים בהשפעתו על החיזוי.

ב. לכן, כדי לשמור על משקל זהה בחשיבות כל פיצ׳ר בתהליך הלמידה, ביצענו נרמול של המשתנים הנומריים לתחום $[0, 1]$ באופן הבא:

$$z_{normalized} = \frac{z - z_{\min}}{z_{\max - z\_min}}$$

כאשר Z הוא הערך שאותו אנו רוצים לנרמל, $z_{min/max}$ הם הערכים המינ׳/המקס׳ בעמודה של Z, בהתאמה.

4. **תרגום הערכים לוקטור *features* כללי**

א. כדי שנוכל להשתמש במודלים הנמצאים בספריית $MLlib$ העברנו את הפיצ׳רים לוקטור.

ב. יצרנו עמודה חדשה בשם $features$ שמחזיקה את הערכים של הרשומה (לאחר עיבוד) בצורה וקטורית.

ג. לבסוף בחרנו את העמודות $features, label\ (gt)$ ואיתן ניגשנו לאימון המודל.

מודל נבחר – *Logistic Regression*

1. חילקנו את הדאטה לסט אימון וסט מבחן, כך שסט האימון מכיל 70% מכלל הדאטה.
2. אימנו את המודל על הדאטה, כפי שנלמד בתרגול.
3. ביצענו חיזוי לסט המבחן וחישבנו את מדד ה-$Accuracy$.
4. מאחר וקיבלנו ערך של כ-70% דיוק, בהתאם לסף הנדרש בפירוט המשימה, לא המשכנו לבדיקת מודלים נוספים.