קובץ זה נכתב לפני סיום הפרויקט והגשתו, על מנת לעשות סדר למי שמעוניינים לעיין בו כבר בשלב הזה. מתוך הנחה שלחבר את הנקודות בצורה מדויקת ניתן לעשות רק כשהפרויקט הושלם, קובץ זה ייתן מענה חלקי לתמונת המצב העדכנית הבפרויקט.

- **<u>מבנה הקבצים והתיקיות</u>** : בתיקיה הראשית יש מילון טבלאות ויפצ'רים שקיבלתי מהמנחה שלי בהתמחות עבור הפרויקט, וקובץ הצהרת כוונות לקראת הכנת הפרויקט, שכתבתי. קבצים אלה ביחד נותנים תמונה די מקיפה על הדאטה ועל מטרות ומבנה הפרויקט. נקודה אחת שהשתנתה משמעותית מאז כתיבת המסמך – בסופו של דבר הפרויקט התמקד רק באחת מ-2 המטרות – חיזוי מגע בין כלים לחלקי מתכת, ולא זיהוי אנומליות. זאת בשל עיכוב בהעברת הדאטה הרלוונטית מהלקוח.
בתיקיה הפנימית, תחת הכותרת "eda", ישנו כל הקוד של שלב זה בתוכנית, דוחות שהופקו, וגם קובץ הסבר שמלווה את החלק הראשון של העבודה.חלק מהדאטה ומהקבצים שנשמרו תוך כדי עבודה לצרכי גיבוי, אשר מלכתחילה היו בתיקיה זו, לא הועלו לגיטהאב. זאת עקב הודעת שגיאה של גיטהאב הקשורה בגודלם של הקבצים.

- **<u>מחברות קוד ושלבי העבודה</u>** : מלכתחילה, התייחסנו לשלב הEDA, לצד הכנת הדאטה והרצת מודל ראשוני לצרכי בדיקה, כשלב אחד. המחברת "EDA and Baseline Model.ipynb" (ולצידה גם המחברת הקטנה "mechkar.ipynb") וקובץ ההסבר – "EDA & Baseline Model" (docx או pdf) מהווים ביחד את ניסיוני הראשון לביצוע שלב זה. יש לציין שלקראת שלב זה לא קיבלתי כמעט הוראות ספציפיות, ועשיתי אותו על פי ניסיוני מהקורס.
לאחר הגשת קבצים אלו, קיבלתי משוב ובו הוראות נוספות, הקשורות בהכנת הדאטה. הביצוע שלהן הינו במחברת "preprocession and baseline model.ipynb". יש לציין שלמחברת זו לא כתבתי עדיין קובץ הסבר, כך שעיקר ההסברים לעבודתי במחברת זו, יוצגו בקובץ זה, להלן.

- **<u>הכנת הדאטה במחברת "preprocession and baseline model.ipynb":</u>**
במסגרת ההכנה מחדש של הדאטה, קיבלתי שני דגשים מרכזיים :
  - את ההפרדה לסט אימון וסט מבחן, לעשות כאשר כל session info id מתקיים בשלמותו ולא מתפצל בין הסטים. זאת מכיוון שבפרודקשן כך דברים ייראו. כמו כן – כל תהליך של feature enrichment לבצע לכל session info id לחוד.
  - לייצר פיצ'רים המתייחסים לדאטה בתור סדרה עתית.

את ההפרדה בין הסשנים ביצעתי בתאי קוד 9-10 במחברת. במסגרת חיפושיי לדרכים להעשרת דאטה המהווה סדרה עתית, גיליתי את ספריית tsfresh ששימשה אותי למטרה זו. בתא 15 השתמשתי בספריה זו, עדיין לכל סשן איי די לחוד. המהלך יצר 2367 פיצ'רים חדשים, כפי שמשתקף בתא 16.
כעת, עם כמות כה משמעותית של משתנים, ניתנה חשיבות גדולה לתהליך הfeature selection. בין התאים 117-124 הרצתי על סט האימון שלי 4 מתודות של feature importance, במטרה לראות מה המשתנים החשובים על פי מתודות שונות. הרעיון הוא שפיצ'ר ש"נבחר" על ידי מספר מתודות, הוא כנראה פיצ'ר שאני

מעוניין להשתמש בו. השאלה הנשאלת – כמה מתודות צריכות "לבחור" בפיצ'ר, על מנת שאכלול אותו בדאטה הסופית לאימון?

על מנת לענות על שאלה זו, יצרתי מספר דאטה סטים, כאשר כל דאטה סט מייצג תשובה אחרת לשאלה זו. בנוסף, יצרתי דאטה סט שמבוסס רק על המשתנים ש"בחרה" מתודת הlasso, מתוך מחשבה שזו עשויה להיות מתאימה לדאטה שלנו, אשר מאופיינת בקשרים הדוקים יחסית בין המשתנים שלה. עבור כל דאטה סט, הרצתי מודל ראשוני לצרכי בדיקה (baseline model) כדי לבדוק באיזה מהם הביצועים הטובים ביותר. תהליך זה נעשה בין התאים 180-228.

כעת (תא 228) ניכרו הפרשים לא גדולים בין התוצאות, כאשר דווקא הדאטה סטים בעלי כמות הפיצ'רים הקטנה ביותר, הביאו לתוצאות הטובות ביותר. המחשבה שלי היא שכיוון שההפרשים בתוצאות קטנים, וההפרשים בין כמות הפיצ'רים היא גדולה, ייתכן ששווה לבחור לא במודל עם התוצאות הטובות ביותר, אלא במודל עם תוצאות טובות מעט פחות אבל עם יותר פיצ'רים. זאת מתוך המחשבה שמודל כזה יהיה כנראה יציב יותר. פניתי עם הדילמה למנחה שלי, ומה שהוא ביקש זה לבדוק עבור כל אחד מהדאטה סטים, עד כמה התוצאות אחידות לפי session info id. כלומר : לבדוק האם עבור כל דאטה סט יש סשנים המועדים לריבוי שגיאות, כאשר מבחינתנו דאטה סט טוב הוא דאטה בו השגיאות מתחלקות באופן שוויוני יחסית בין הסשנים.

בשלב הזה, של מענה להנחיה זו, הפרויקט נמצא כרגע. בין התאים 300-353 אני מבצע את הבדיקה עבור דאטה סט אחד (על בסיס lasso), ובתא 394 אני כותב פונקציה שמגדירה את התהליך, על מנת שאבצע אותו דבר לדאטה סטים הנוספים. אני מריץ את הפונקציה על הדאטה סטים הנוספים בתאים הבאים.

ניתן לראות שהגדרתי כמה דרכים לבחון את סוגיית התחלקות השגיאות בין הסשנים – הפרש ממוצע השגיאות לעומת ממוצע השגיאות הכללי, שגיאה מקסימלית לעומת ממוצע השגיאות הכללי, וכמות פערים "גדולים", של למעלה מ0.2 לעומת ממוצע השגיאות הכללי. מדובר במדדים ראשוניים, ובוודאי ניתן ליצור נוספים, אבל זו הנקודה בה עומד הפרויקט כרגע.

- **<u>הצעדים הבאים</u>** : כעת שוב אני עומד בפני החלטה על בחירת הדאטה סט הסופי עליו אאמן מודלים. ושוב, כל המדדים מצביעים לדאטה סט הקטן, בעל 12 הפיצ'רים. לאחר התייעצות, אבחר דאטה סט, ואז אאמן עליו מודלים. המודל שיקבל את התוצאות הטובות ביותר, ייבחר לשלב הבא, של אופטימיזציה למודל. לאחר אופטימיזציה אבדוק את המודל על הtest, אצור קובץ מודל ואשמור את התהליך באופן שייאפשר deployment.

- **<u>הבהרות קטנות נוספות:</u>**

- שם הלקוח עבורו נעשה הפרויקט, כמו גם פרטים אחרים, נמחקו בכוונה מכל הקבצים בתיקיה פומבית זו. מדובר בפרטים שהינם תחת חוזה סודיות.
- ניתן לזהות לא מעט חזרתיות על קוד בכתיבה שלי ומעט מידי שימוש בכתיבת פונקציות. להערכתי לקראת סוף העבודה ניכר השיפור בתחום. בכל מקרה הפרויקט כפי שיוגש יהיה לאחר סדרת תיקונים בנושא.

- יש להניח כי ישנן עוד נקודות שהיה ראוי להעלות על הכתב בקובץ הסבר זה, ולא עלו על דעתי בנקודת זמן זו ושלב זה של העבודה. אשמח לתת מענה לכל שאלה הקשורה הפרויקט, בדאטה, באופן החשיבה וכיוצא בכך.