תקציר

פרויקט זה נכתב ב-2 מסגרות :

1. כפרויקט גמר בתוכנית הכשרה של מדעני ומתנחי נתונים באוניברסיטת בר אילן.
2. כחלק מהתמחות בחברת היעוץ Modernize Solutions, במסגרת עבודה שנעשתה עבור לקוח של החברה.

תיקיה זו הינה הגרסה הפומבית של העבודה, המשמשת עבור פרויקט הגמר, ומספר נקודות הקשורות בסודיות הושמטו ממנה. לאורך הדרך ניווטתי את הפרויקט וההחלטות במהלכו בהתאם לדרישות פרויקט הגמר, אך כשעלו צרכים שונים מהצוות בModernize Solutions ו/או מהלקוח, חרגתי מההנחיות המקוריות.

הלקוח שעבורו נעשתה העבודה הינו חברה שמייצרת צ'יפים לבתי חרושת שמייצרים חלקי חילוף מתכתיים לספינות, מטוסים ועוד. אותם צ'יפים כוללים חיישנים שמשדרים מידע בדבר תהליך היצור במכונות השונות לאורך זמן, ובתוך כך בעיקר נתונים אודות החום, הרעש ותדירות התנועה במכונות ובכלים השונים.

הדאטה

כלל הדאטה הרלוונטית, ובתוך כך הנתונים מהחיישנים, מרוכזים בשירות הענן SnowFlake, ב-13 טבלאות. פירוט של הטבלאות ישנו בקובץ "ERD.png" בתיקיה זו. עם זאת, לצורך הפרויקט הנוכחי, היה צורך בשימוש ב-3 מתוך הטבלאות:

1. Session_data – בה מרוכזים הנתונים מהחיישנים : החום, הרעש ותדירות התנועה במכונות ובכלים השונים, לפי יחידות זמן של דגימה אחת למילי שניה.
2. Session_info – בה מרוכז המידע על כל אחת מהמכונות במפעל, ובתוך כך מידע על סוג הכלי של כל מכונה, רמת טעינה הסוללה שלו, ועוד. כיוון שכל אחת מהדגימות הקשורות לטבלה הקודמת, קשורה למכונה ספציפית, הטבלאות מקושרות.
3. Labels – בטבלה זו אנו מגלים עבור כל דגימה מSession_data , האם יש בה מגע בין כלי לבין חלקי מתכת, והאם זוהתה בה אנומליה. כפי שניווכח עוד זמן קצר, נושא המגע בין הכלי לחלקי המתכת זו הנקודה אותה נתבקשנו לחזות. בנוסף – יש בטבלה זו מספר טורים אשר תויגו לאחר מעשה, וקשורים בכל אחת מהדגימות. הנתונים מטורים אלה לא יכולים לשמש לפרידיקציה, כיוון שזה מידע שנאסף בדיעבד, וגם בעתיד לא יהיה זמין בזמן אמת עבור הפרידיקציה. בחלק מהמקרים מדובר בנתונים שהם פועל יוצא של הפרידיקציה, כלומר שימוש בנתון היה מהווה את הנחת המבוקש.

פירוט רחב יותר של המשתנים השונים של שלושת הטבלאות, ניתן למצוא בקובץ "Data_Dictionary.xlsx" שבתיקיה זו.

המשימה

המטרה שלנו היא לחזות עבור כל אחת מהדגימות האם מתקיים בה מגע בין הכלי לבין חלק מחלקי המתכת. מלכתחילה היה רצון לבצע פרידקיציה גם עבור שאלת האנומליות, אך בשלב זה מטרה זו ירדה מהפרק עקב מחסור בנתונים. המגע בין הכלי לחלקי המתכת משמעותי מאוד ללקוח, כיוון שהוא מהווה בסיס לחישובים נוספים שלהם הקשורים בתהליך היצור, בין היתר חישובים הקשורים לבלאי של המכונות.

מבנה הקבצים והתיקיות בפרויקט

כפי שצוין, בתיקיה זו (התיקיה הראשית של הפרויקט) מלבד קובץ זה ישנם שני קבצים הקשורים להבנת הדאטה:

1. ERD.png מפרט את הטבלאות, אילו פיצ'רים יש בכל טבלה, והקשר בין הטבלאות.

2. Data_Dictionary.xlsx - מילון פיצ'רים; מסביר את הפיצ'רים של 3 הטבלאות הרלוונטיות לפרוויקט.

מבחינת תיקיות המשנה תחת תיקיה זו, ישנן מספר תיקיות שאינן רלוונטיות למעיין בעבודה:

1. Venv – תיקיית סביבה וירטואלית, בה שמורות ההתקנות השונות.
2. .git – תיקיה אוטומטית שנוצרה עם הגדרת הפרויקט כפרויקט בGithub. משמשת לצרכים של Github כגון סינכרון.
3. .idea – כנ"ל.

שתי תיקיות פנימיות שאחריהן כן כדאי לעקוב, ובהן שמורים הקוד והדאטה של הפרויקט:

1. eda, preprocession, baseline model – תיקיה זו, כשמה כן היא, והיא כוללת מחברות קוד, דאטה ודוחות הקשורים לנושאים הכלולים בשמה :
   - eda
   - preprocession
   - baseline model

2. training and select models– גם תיקיה זו, כשמה כן היא, עוסקת באימון מודלים ובבחירת המודל הסופי לפרוויקט.

ב-2 התיקיות הללו, 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. עם זאת, את התהליך שעברתי עם הקבצים הללו ניתן לראות במחברות הקוד.

<u>מחברות קוד ושלבי עבודה</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") מהווים ביחד את ניסיוני הראשון לביצוע שלב זה. יש לציין שלקראת שלב זה לא קיבלתי כמעט הוראות ספציפיות, ועשיתי אותו על פי ניסיוני מהקורס.
mechkar.ipynb תורמת בכך שדרכה הפקתי דוח אוטומטי של חבילת "Mechkar" בשפת R.

לאחר הגשת קבצים אלו, קיבלתי משוב ובו הוראות נוספות, הקשורות בהכנת הדאטה. הביצוע שלהן הינו במחברת "preprocession and baseline model.ipynb" ובה גם פירוט על השינויים שנתבקשתי לבצע.

בשלב זה הדאטה הייתה מוכנה ונותר היה לאמן ולבחור מודלים. זאת נעשה בתיקיה הרלוונטית, בקובץ "model selection.ipynb".

<u>הצעדים הבאים</u>

המועד האחרון להגשת הפרויקט הגיע, ועל פניו יש בידיי פרויקט יפה עם ניתוח מורכב ותוצאות מצויינות. עם זאת, עדיין יש מה לשפר, ומספר צעדים צפויים במסגרת ההתמחות, ואלו בתקווה יבוצעו ויעודכנו גם כן בGithub בקרוב:

- הרצת מודל LSTM לדאטה. כיוון שמדובר במודל מוצלח לסדרות עתיות, יש מחשבה שאולי הוא יוכל להתאים יותר ממודלים קלאסיים של למידת מכונה. הדבר מצריך בדיקה ולימוד נפרדים, ובתקווה יעשיר את הפרויקט בדיעבד.
- בניית פונקציות שיטמיעו את הפרמטרים של המודל בצ'יפים של הלקוח, באופן שיאפשר חיזוי בזמן אמת.
- בניית מודל גם עבור אנומליות בתהליך היצור, כשיהיה מספיק דאטה עבור כך.