פרויקט סיום בנושא סיווג (CLASSIFICATION)| מבוא לבינה מלאכותית 10101

מציג: שובל בנזר ת.ז 319037404

מבוא

"Perfection does not exist but you have to look for it anyway" הוא משפט של מאמן כדורגל ענק - פפ גווארדיולה בתמונה. תמיד אהבתי כדורגל. מעולם לא הייתי טוב בכך, ההיפך הוא הנכון. מהמקום שמאוד אהבתי את המשחק עם הכדור העגול אך לא הייתי טוב בו, התפתחה אהבה בלהבין את המשחק. בהתחלה כאוהד ביציע, ובתיכון ביליתי לילות רבים בשני משחקים במיוחד – football manager, fifa. בראשון הצלחתי לשלוט על 11 השחקנים ולגשר על הפער כאשר אני עצמי במגרש, והשני בלנסות לשלוט על המועדון מרמה מעט גבוהה יותר. באותם משחקים, הערכת השחקנים נתפסה כמשימה יחסית פשוטה, על פי קריטריונים ספציפיים שקבעה החברה המפתחת. עם זאת, במציאות, קיימת בעיה מאחר ומדידת נתונים כגון נחישות, מקצוענות, נטייה לפציעות וגנטיקה היא מורכבת וקשה לקליטה. המעבר מחיי המשחק למציאות עצמה של משחק כדורגל אמיתי, בו עליי לקחת בחשבון את התנהגותם של 22 שחקנים תחת חוקי פיזיקה מוחשיים, מסבך את הרצון להבין אפילו יותר. במסגרת פרויקט זה, אבקש לחקור את האפשרות לחזות את שווי של שחקן כדורגל בהתבסס על נתונים נגישים לקהל רחב, באמצעות שימוש באלגוריתם הפרספטרון.

DATASET ACQUISITION

אתר טרנספרמרקט הוא אתר עולמי חינמי למעקב אחר נתוני שחקני כדורגל, והצלחתי למצוא dataset מ09.04.24 כלומר נתונים עדכניים לעבוד עליהם.

הורדתי שתי טבלאות המהוות מסד נתונים רלציוני, וניתן לקשר באמצעות מזהה שחקן בין הטבלה הראשונה לשנייה:

א. טבלת שווי שחקן (מזהה שחקן, שווי שחקן).

ב. טבלת הופעות של שחקן (* (*מזהה משחק, מזהה שחקן, שם שחקן, דקות במשחק, שערים במשחק, בישולים במשחק, צהובים במשחק, אדומים במשחק)

DATA PREPROCCESSING

בטבלת הופעות של השחקן, לכל מזהה שחקן קיימים מספר ערכים לפי מזהה משחק. כדי לנהל ולנתח את המידע, בחרתי לקבץ את הערכים ולחשב עבורם ממוצע, חציון וסטיית תקן. הערכים הנומריים שנמדדים הם כרטיסים צהובים, כרטיסים אדומים, בישולים, שערים, ודקות ששוחקו.

לאחר ההורדה של הדטהסט שבחרתי, גודל הטבלה הופעות עמד כ-1.56 מיליון שורות על 7 עמודות. לאחר קיבוץ הערכים לפי הערכיים הנומריים שבחרתי והסרת הערכים כפולים, מספר העמודות עלה ל-22, והמספר הכולל של השורות בנתונים הוא 19,105.

לאחר שלב הזה הפלט של הטבלה נראה כך (5 השורות הראשונות ו9 מהעמודות הראשונות לדוגמה):

	player_id	yellow_cards	red_cards	goals	assists	minutes_played	market_value	yellow_cards_mean	yellow_cards_std
0	38004	0	0	2	0	90	NaN	0.030303	0.172733
1	79232	0	0	0	0	90	NaN	0.100000	0.316228
2	42792	0	0	0	0	45	NaN	0.125000	0.353553
3	73333	0	0	0	0	90	NaN	0.076923	0.271746
4	122011	0	0	0	1	90	NaN	0.090909	0.288355

כעת, כפי שניתן לראות בעמודת market_value נדרשתי להוסיף ערכים מטבלת שווי השחקן לטבלת הופעות שערכתי לאחר שקיבצתי את הופעות השחקנים לערכים ייחודים לכל שחקן. בעת ניסיון להוסיף שווי שוק לכל שחקן, נתקלתי בשגיאות קוד שלא הצלחתי לפתור בקלות. לדוגמה, בניסיון להוסיף את ערכי שווי השחקן לטבלה, נתקלתי בכפילויות של שחקנים לאחר שכבר הורדתי כפילויות, מה שדרש ממני להתחיל את התהליך מחדש רק שהפעם אוסיף את השווי שוק לשחקן לפני שאקבץ אותם – מה שיכול להאריך משמעותית זמן ריצה לנתונים גדולים יותר. מבחינת זמן עבודה – לתקן שגיאות קוד ולוודא שהפלט שקיבלתי מתאים לדטהסט שרציתי לבנות, היה אתגר משמעותי מאוד בפרויקט זה.

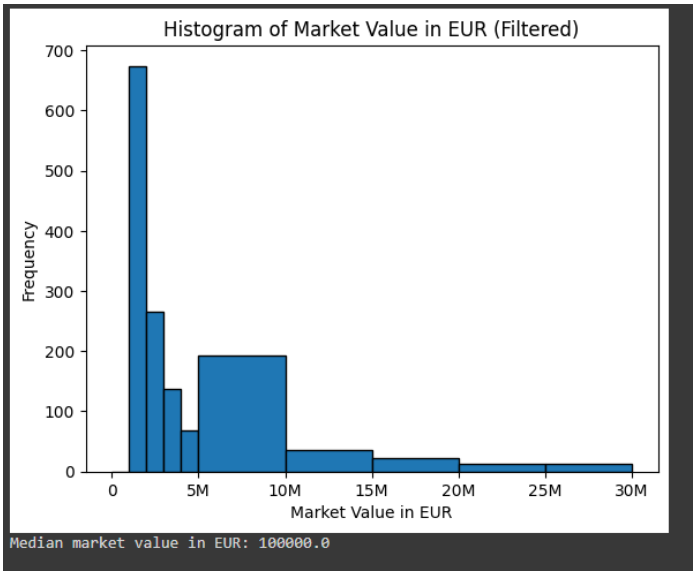
MODEL ARCHITECTURE

MODEL SELECTION

בחרתי באלגוריתם הפרספטרון כסוג של מסווג לינארי בפרויקט שלי. הסיבה לבחירה בפרספטרון כוללת את פשטותו ויעילותו בלמידת מפרידים לינאריים. היכולת שלו לעדכן את משקלותיו באופן איטרטיבי הופכת אותו למתאים עבור מערכי נתונים גדולים כמו זה שהתמודדתי איתם.

DATA PREPARATION

סיננתי את ערכת הנתונים כך שתכלול רק מופעים שבהם הערך שווי שוק באירו הוא לפחות 100,000 אירו, תהליך שמהווה צורת נירמול של המאפיין כפי שניתן לראות בגרף המצורף. לאחר מכן, יצרתי יעד לסיווג בינארי, שבו ערך שווי שוק מעל 1,000,000 אירו נחשב למחלקה חיובית (1), ומתחת או שווה לכך נחשב למחלקה שלילית (0). הסרתי את המאפיין market_value_in_eur מרשימת המאפיינים שלי, כדי למנוע דליפה מכיוון שהוא קשור ישירות לתווית.



DATA SPLITTING

הפרדתי את הנתונים לשני סטים: אימון ובדיקה, וסיננתי אותם כך שיכללו רק ערכי שוק הגבוהים ממאה אלף יורו. בחרתי בסינון זה כדי להתמקד בתת-קבוצה של הנתונים כדי לקבל התפלגות יותר ממוקדת עבור התופעה שאני מנסה לחזות, שחקנים עם שווי שוק. כמו כן, חילקתי את הנתונים לסט אימון בגודל של 80% וסט בדיקה בגודל של 20% בצורה רנדומית בעזרת הפונקציה test_train_split מהספריה sklearn.

MODEL TRAINING

המודל התאמן באמצעות פונקציית הגרדיאנט הסטוכסטי (SGD) שלמדנו בכיתה, תוך שימוש בפונקציה fit גם מהספריה sklearn היא מיישמת את אלגוריתם הפרספטרון. מאחר ומדובר בבסיס נתונים גדול זמן החישוב שלה עדיף, אם זאת GD על כל הבסיס נתונים יכול לתת תמונה מדויקת יותר. השיקול שבחרתי בו הוא זמן חישוב על מנת שהמודל יהיה יעיל גם מול בסיסי נתונים גדולים יותר.

MODEL EVALUATION

בדקתי את ביצועי המודל שלי באמצעות מדדי הערכה: דיוק (Accuarcy), נכונות(Precision), ריקול (Recall) וציון ROC AUC .

באמצעות דיוק, יכולתי לזהות כמה מהמופעים שסווגו כחיוביים היו באמת חיוביים. ריקול עזר לי לזהות כמה מהמקרים החיוביים האמיתיים זוהו כחיוביים במודל. ובסוף, ציון ה ROC AUC-סיפק לי מידה הסתברותית לדירוג נכון של התחזיות של המודל.

ממדדי ההערכה שהתקבלו, המודל הציג ציון דיוק גבוה, מה שמרמז שהוא סווג נכון חלק גדול מנתוני הבדיקה. עם זאת, יש לפרש את הדיוק הגבוה עם זהירות בשל אי האיזון שמוצע על ידי ציוני הנכונות והריקול.

הנכונות של 75% מצביע על כך שכאשר המודל חוזה ערך שווי שוק מעל מיליון אירו, הוא נכון 75% מהזמן. עם זאת, מדד זה לבדו אינו נותן חשבון על מספר המקרים החיוביים האמיתיים בנתונים.

ציון הריקול נמוך מאוד, מה שמצביע על כך שהמודל זיהה רק כ-1.35% מערכי השוק האמיתיים מעל מיליון אירו בסט הבדיקה. זה מרמז שלמרות שהמודל מדויק בתחזיותיו החיוביות, הוא מפספס מספר גדול של מופעים חיוביים אמיתיים.

ציון ה ROC AUC-נמצא סביב 0.57, שהוא רק מעט טוב יותר מניחוש אקראי (שיהיה לו ציון של 0.5). ציון זה לוקח בחשבון גם את שיעור החיוביים האמיתיים וגם את שיעור החיוביים השגויים ומצביע על כך שלמודל אין יכולת הבחנה חזקה להבדיל בין המחלקות.

מסקנה :התוצאות מרמזות שלמרות שהמודל יעיל בהימנעות מחיוביים שגויים (כפי שמראה הדיוק), הוא נכשל בלתפוס את רוב החיוביים האמיתיים (כפי שמראה הריקול). הציון הנמוך של הריקול עשוי להוביל למסקנה שיש מספר גבוה של שליליים שגויים, שבהם מקרים בעלי ערך שווי שוק גבוה אינם מזוהים.

```
(19105, 22)
Accuracy: 0.9424234493588066
Precision: 0.75
Recall: 0.013513513513513514
ROC AUC Score: 0.5066178292769014
```

.

APPLICATION TO THE MULTI-LABLE CASE

המודל מקטלג ערכי שוק באמצעות תנאים מדורגים: מעל 2 מיליון אירו כתווית 4, מעל מיליון אירו כ-3, מעל 500,000 אירו כ-2, ומעל 100,000 אירו כ-1. מתחת למאה אלף היא תווית 0. גישה מדורגת זו נלקחה כדי לתפוס את הרמות בערך השוק, במטרה לשפר את העדינות של החיזויים והרלוונטיות לשונות שווקים.

הדוגמא של מודל הסיווג הרב-תוויות מציגה דיוק צנוע של בערך 0.557, מה שמצביע על ביצועים טובים מעט מסיכוי אקראי. עם זאת, הדיוק הממוצע לפי מקרו נמוך ועומד על 0.111, מה שמגלה את האתגרים בחיזוי נכון של חיוביים אמיתיים בכל המחלקות. ריקול ממוצע לפי מקרו של 0.2 מצביע על כך שהמודל מצליח לזהות רק

חמישית מכל המופעים הרלוונטיים. הממוצעים המיקרו והמשוקללים לדיוק ולריקול נשפעים יותר מהמחלקות הדומיננטיות בשל אי האיזון המחלקות.

על מנת לשפר את המודל בעתיד, יש להתמקד בשיפור בחירת התכונות כדי לתפוס טוב יותר את הקשרים המורכבים בין התוויות, לטפל באי האיזון בין המחלקות הנמדדות, שימוש במודלים מתוחכמים יותר, לכוון את פרמטרי המודל ולהשתמש במדדים כמו ציון F1 להערכה מאוזנת. שיקול של הקורלציות בין התוויות והפעלת טכניקות מותאמות לנתונים בסיווג רב-תוויות יכולות גם לעזור בשיפור הביצועים החיזויים של המודל .

```
Accuracy: 0.5574456948442816
Macro-average Precision: 0.11148913896885633
Macro-average Recall: 0.2
Micro-average Precision: 0.5574456948442816
Micro-average Recall: 0.5574456948442816
Weighted-average Precision: 0.31074570270042395
Weighted-average Recall: 0.5574456948442816
```

קישור לקוד (מחברת בjupyter)