מסמך מסכם

פרויקט בפינטק תיאור המשתנים והנכס לחיזוי:

מגישים: רביד גרש, יובל מרגולין, רן סלע וירדן נחום

הנכס לחיזוי הוא שווי של שחקן כדורגל שנה ושנתיים קדימה. השווי נקבע על פי אנליסטים של הארגון FIFA, השווי נמדד כל שנה ומפורסם בחודש ספטמבר.
מספר הפיצ'רים שבחרנו הוא 93. בחירת הפיצ'רים התבססה על פי פיצ'רים מרכזיים ש-FIFA מייחסת להם חשיבות בהערכת שווי שחקן ודירוגו, וכן הפיצ'רים הללו רלוונטיים במאמרים שקראנו שמנסים לחזות את שווי השחקן.
דוגמה לפיצ'רים מרכזיים שבחרנו : נתונים פיזיים כדוגמת גיל, גובה ומשקל, שכר נוכחי של השחקן ושווי נוכחי וגם ציונים במדדים שונים של מהירות ויכולות משחק שונות שנותחו לפי אנליסטים עבור FIFA.

בנוסף לפיצ'רים הללו יצרנו איברי אינטראקציה בין פיצ'רים מסוימים מתוך 93 הפיצ'רים. כלומר מכפלה של איברים וזאת על סמך קשר חזק ביניהם שהתגלה לפי המאמרים שסקרנו (למשל ככל ששחקן מבוגר יותר ונתון פיזי כדוגמת ריצה נמוך יותר אנחנו צופים שהשווי שלו יהיה פחות מאשר שחקן צעיר עם נתון גבוה מתוך הנחה שהצעיר יוכל להשתפר או לכל הפחות לשמור על הרמה הגבוהה לאורך זמן. בנוסף ביצענו גם העלאה של כלל המשתנים בריבוע כדי לבדוק קשר פולינומי מדרגה שנייה של הפיצ'רים לשווי העתידי של שחקן. יחד עם איברי האינטראקציה יצרנו איברים נוספים שמשלבים את הפיצ'רים הקיימים למשל פוטנציאל פחות יכולת כללית נוכחית כדי להשתפר ביכולת שיפור עתידית של השחקן בתור פי'צר.
סך הכל היו לנו כ-280 פיצ'רים אחרי שהוספנו את איברי האינטראקציה אבל במודלים שלנו בחרנו רק חלק מתוכם, בפרט את אלו שהיו בעלי ההשפעה הגדולה ביותר לקביעת שווי השחקן.

מקורות הנתונים:
מקור הנתונים שלנו הינו FIFA 24 Complete player dataset הכולל נתונים על כ-80,000 שחקני שחקנים משנת 2015 עד 2023. לכל אחד מהשחקנים יש לנו ציון בפרמטרים השונים בשנה כלשהי ואת השווי של אותו השחקן בשנה ושנתיים לאחר מכן. במקור קובץ הנתונים מכיל 109 פיצ'רים לכל שחקן (אנחנו בחרנו לסנן חלק מהם נפרט על כך בהמשך), הפיצ'רים השונים כוללים סטטיסטיקות שונות על השחקן כמו יכולת תקיפה, כישורים, הגנה, מנטליות, מהירות, העמדה שבה משחק השחקן וגם המעדון שבו הוא משחק. כמו כן, מכיל מידע אישי אודות השחקן כמו עיר הולדתו, שנת הלידה והשכר שלו.
הדאטה נלקח מהאתר Kaggle, קישור לנתונים : FIFA Data.

לאחר שניקינו את הנתונים נשארו בידינו:

כמות הרשומות: 37,346 שחקנים מתוכם - 3,662 שוערים, 7,091 בלמים, 6,474 מגנים, 7,871 קשרים הגנתיים, 7,821 קשרים התקפים, 4,427 חלוצים.

<u>שיטות לניקוי וסנכרון הנתונים</u>

ראשית, מאחר שהשתמשנו בנתונים של FIFA ממספר שנים, היינו צריכים לבצע התאמות בין שמות השחקנים מאחר שהשמות לא תמיד היו עקביים (למשל חלק מהנתונים כללו גם את השם האמצעי של השחקן). לפיכך, בנינו פונקציה שעושה ניקוי לנתונים ומעדכנת אותם, על מנת שנוכל בסופו של דבר למזג בין השנים השונות שיש לנו בדאטה.
כמו כן, ביצענו שינויים בשמות העמודות כדי שיהיה לנו נוח יותר לעבוד איתן. בנוסף, הורדנו מספר לא מבוטל של פיצ׳רים שהיו לא רלוונטים עבור המשימה שלנו או שהיו כפולים בדאטה.
<u>דוגמאות לפיצ׳רים שבחרנו לא להשתמש בהם</u>: המזהה הייחודי של השחקן, מספר החולצה שלו, תמונה של דגל המועדון ומדינה, תמונת פנים של השחקן ועוד.

בנוסף על סמך המאמרים שבדקנו, ביצענו שינויים נוספים בדאטה והפכנו עמודות שאינן מספריות למספריות. כך למשל הפכנו את עמודת העמדה של השחקן למספרים בין 1-6 (1 עבור השוערים, 2 עבור בלמים, 3 עבור מגנים, 4 קשרים אחוריים, 5 קשרים התקפיים ו-6 עבור החלוצים).

כמו כן, לכל שחקן הייתה עמודה שפירטה על התכונות המיוחדות שלו, למשל ״בעיטה מסובבת״, ״חסין לפציעות״ וכו׳. בחרנו ליצור עמודה חדשה שסוכמת את כמות התכונות המיוחדות של השחק מתוך תפיסה שכמות התכונות הייחודיות מעידה על איכותו של השחקן.

בנוסף, ביצענו המרות של משתנים קטגוריים לערך מספרי. למשל, ישנה עמודה בשם WORK RATE והאפשריות בה הן קטגוריות: high, low וכו׳. הפכנו אותם למספרים בסדר מדורג, המספר הכי גבוה ניתן ל-HIGH.
בנוסף לפי דירוג הקבוצה על פי FIFA חילקנו את הקבוצות ל-6 רמות (רמה 1 - הקבוצות בעלות דירוג 5 כוכבים ולאחר מכן יורד 4.5,4 וכו׳), באופן דומה חילקנו את הנבחרות ל-5 קבוצות.
יש לציין שישנן מעט קבוצות ברמה 6 ועבורן הנתונים היו פחות מדוייקים בשל חוסר הפופולריות של הקבוצות ולכן עבור תחזית של שנתיים קדימה לא נעזרנו בנתונים של שחקנים שלהן (אולם עבור שנה קדימה כן נעזרנו).

היו ערכים חסרים שבחרנו למלא אותם באפסים, כאשר היה היגיון בדבר. למשל, יש עמודה שמעריכה את היכולת של השחקן בתור שוער (נותנת לו ציון מספרי). עבור שחקנים שהם לא שוערים היה שם NAN, ולכן מצאנו למלא את הערך הזה ב-0 שכן הוא לא רלוונטי לשחקן שהוא לא שוער ולא נרצה שהמודל שלנו יתן לערך הזה משקל מסוים.

<u>מודלים</u>

המודל הראשון שבחרנו לבדוק הוא מודל **רגרסיה לינארית** לכל אחת משש קבוצות השחקנים,
כלומר עבור בלמים בנינו רגרסיה שונה מאשר הרגרסיה שבנינו עבור חלוצים (כי הפיצ׳רים שרלוונטיים אליהם שונים).
בחרנו במודל זה בגלל מספר סיבות שונות. ראשית, שימוש ברגרסיה לינארית לחיזוי שווי עתידי של שחקנים

מציע מספר יתרונות בולטים. המודל מאפשר לזהות קשרים והסתמכויות ישירות בין תכונות השחקן, כמו גיל, ניסיון וקבוצה נוכחית, לבין שווי השוק שלו, תוך הבנה ברורה של

תרומת כל תכונה לשווי. בנוסף, המודל קל יחסית למימוש ולפרשנות, מה שמאפשר להסיק מסקנות ברורות לגבי השפעת התכונות השונות על השווי. כך ניתן לאפיין את תרומת כל פיצ׳ר ולהשתמש במודל כדי לבנות תחזיות.

<u>נציג את התוצאות הטובות ביותר שהתקבלו כאשר מספר הפיצ׳רים הוא 9:</u>

	Position	Num Features	Prediction Ahead	R^2	MSE	Top Features
0	GK	9	1 Year	0.864916	13.130007	['value_eur', 'age_movement_reactions ** 2', 'age_movement_reactions', 'age_club_rank ** 2', 'age_potential ** 2', 'gk ** 2', 'movement_reactions ** 2', 'age_club_rank', 'age_skill_dribbling ** 2']
1	GK	9	2 Years	0.739528	63.347250	['value_eur', 'age_movement_reactions ** 2', 'age_movement_reactions', 'age_club_rank ** 2', 'age_skill_dribbling ** 2', 'age_potential ** 2', 'overall ** 2', 'age_international_reputation ** 2', 'international_reputation']
2	CB	9	1 Year	0.881033	8.120482	['value_eur', 'age_defending_sliding_tackle ** 2', 'age_defending_sliding_tackle', 'age_club_rank', 'age_potential ** 2', 'club_rank ** 2', 'club_rank', 'age_movement_reactions ** 2', 'gk']
3	CB	9	2 Years	0.652297	44.549965	['value_eur', 'age_defence ** 2', 'age_defence', 'age_dribbling ** 2', 'age_international_reputation ** 2', 'year_contract ** 2', 'age_international_reputation', 'club_rank ** 2', 'movement_agility ** 2']
4	RB LB	9	1 Year	0.726423	12.851102	['value_eur', 'age_passing ** 2', 'age_passing', 'wage_eur', 'year_contract ** 2', 'club_contract_valid_until_year ** 2', 'age_power_stamina', 'age_club_rank', 'release_clause_eur']
5	RB LB	9	2 Years	0.671776	34.911333	['value_eur', 'age_defending ** 2', 'age_club_rank', 'age_movement_agility ** 2', 'club_rank ** 2', 'defending', 'release_clause_eur', 'wage_eur', 'age_attacking_crossing ** 2']
6	DEFENSIVE MID	9	1 Year	0.886786	14.095768	['value_eur', 'age_passing ** 2', 'age_passing', 'age_defence ** 2', 'player_tags ** 2', 'age_club_rank', 'age_potential ** 2', 'club_rank ** 2', 'age_attacking_crossing ** 2']
7	DEFENSIVE MID	9	2 Years	0.808948	30.938143	['value_eur', 'age_skill_long_passing ** 2', 'age_skill_long_passing', 'player_tags ** 2', 'age_club_rank', 'age_potential ** 2', 'club_rank ** 2', 'wage_eur', 'potential']
8	ATTACKING MID	9	1 Year	0.887539	15.298068	['value_eur', 'age_international_reputation ** 2', 'age_skill_dribbling ** 2', 'age_international_reputation', 'age_club_rank', 'age_dribbling ** 2', 'club_contract_valid_until_year', 'player_tags ** 2', 'value_eur ** 2']
9	ATTACKING MID	9	2 Years	0.706259	72.423856	['value_eur', 'age_attacking_short_passing ** 2', 'age_club_rank', 'age_international_reputation ** 2', 'age_attacking_short_passing', 'age_dribbling', 'club_rank ** 2', 'fifa_age ** 2', 'lw ** 2']
10	STRIKER	9	1 Year	0.844551	20.301334	['value_eur', 'age_skill_ball_control ** 2', 'age_skill_ball_control', 'age_club_rank', 'age_international_reputation ** 2', 'international_reputation ** 2', 'club_rank ** 2', 'fifa_age ** 2', 'skill_ball_control']
11	STRIKER	9	2 Years	0.774142	72.233438	['value_eur', 'age_power_long_shots ** 2', 'age_international_reputation ** 2', 'age_power_long_shots', 'fifa_version ** 2', 'player_tags ** 2', 'age_power_jumping ** 2', 'club_contract_valid_until_year ** 2', 'height_cm ** 2']

מודל נוסף שבחרנו לבדוק אותו עבור המשימה שלנו הוא **רגרסיית לאסו** (Lasso Regression).
זו שיטת רגרסיה שמבצעת בחירה של פיצ׳רים והקטנה שלהם במידת הצורך כדי למזער את מספר התכונות ולהתמקד בתכונות המשמעותיות ביותר. בניגוד לרגרסיה לינארית רגילה, לאסו מוסיפה סוג של עונש (רגולריזציה) שדוחף את ערך המקדמים של חלק מהתכונות לאפס, וכך למעשה מסננת את התכונות הפחות חשובות. **<u>התוצאות שהתקבלו מפורטות במסמך המפורט.</u>**

<u>מודל נוסף שבדקנו הוא **Random Forest**</u>

Random Forest הוא מודל למידת מכונה המורכב ממספר עצי החלטה, אשר משלב את התחזיות של כל עץ ליצירת חיזוי יציב יותר. אחד היתרונות המרכזיים בשימוש ב-Random Forest במשימה שלנו הוא יכולתו למנוע Overfitting. המודל בוחר תת-קבוצות רנדומליות של דוגמאות ותכונות עבור כל עץ, מה שמונע מהעצים ללמוד דפוסים ספציפיים מדי בנתוני האימון, וכך מתקבל מודל כללי יותר שמתאים גם לנתונים חדשים. בנוסף, Random Forest מתמודד היטב עם מספר רב של פיצ׳רים, כמו במשימה שלנו שבה יש קרוב ל-300 פיצ׳רים שונים.

באמצעות מנגנון זה, המודל מספק מידע על דירוג חשיבות הפיצ׳רים, ומאפשר לנו למקד את המודל בפיצ׳רים הדומיננטיים ביותר ולבנות מודל פשוט ויעיל יותר.

יתרון נוסף הוא הגמישות של המודל, המאפשרת שליטה במספר הפרמטרים כגון מספר העצים ביער (n_estimators), מספר הפיצ׳רים שנבחרים בכל פיצול (max_features) ופרמטרים נוספים שניתן לכוונן בהתאם למשימה. בנוסף, Random Forest ידוע ביכולת ההתמודדות שלו עם רעש וערכים חסרים בנתונים, מה שהופך אותו למתאים במיוחד למשימה שלנו. השימוש ב-Random Forest מספק איזון בין מניעת אוברפיטינג, יכולת התמודדות עם מספר רב של פיצ׳רים, גמישות התאמת פרמטרים ויכולת לזהות את הפיצ׳רים החשובים ביותר לחיזוי – כל אלה הופכים אותו לבחירה אופטימלית למשימה שלנו.

תוצאות מפורטות עבור מודל Random Forest נמצאות במסמך המפורט.

<u>קישורים למאמרים רלוונטיים</u>

במסגרת העבודה זו, כפי שצוין בשלב האקספלורציה, הסתמכנו על מספר מאמרים שונים.
נעזרנו במאמרים הללו כדי לבחור בין היתר את הפיצ׳רים שלנו, את מקורות הנתונים וכן את המודלים שבחרנו לבדוק.

<u>קישורים:</u>

- What should clubs monitor to predict future value of football players
  ממאמר זה שאבנו השראה בבחירת המודלים שרצינו לבדוק על מנת לחזות את שווי השחקן ובנוסף גם את המטריקות שכדאי להסתכל עליהן כאשר מעריכים שווי של שחקן.

- Poli, Raffaele. "Econometric Approach to Assessing the Transfer Fees and Values of Professional Football Players." Economies, vol. 10, no. 1, 2022, p. 4. MDPI.

- Al-Asad, Mustafa A. "Predict the Value of Football Players Using FIFA Video Game Data and Machine Learning Techniques." IEEE Access, 2022. IEEE.
  מאמר זה מחזק את הרעיון שניתן להסתמך על נתוני FIFA, התבססנו עליו בכך שביצענו שימוש בנתונים של FIFA.

- Calcalist, 2022. Calcalist. "איך נקבע שווי של שחקן כדורגל"
  בכתבה זו עולים מספר פיצ׳רים שנמצאו משפיעים על ערך השחקן על פי מחקרים שמתוארים בכתבה.
- "Explainable Artificial Intelligence Model for Identifying Market Value in Professional Soccer Players." arXiv, 2023. arXiv.

- "The wisdom of crowd, real option and game theory decisions: can they be used by clubs to improve their investment in football players?", Gracia Rubio Martín 2023, Emerald

- Valuing soccer players：on the valuation dynamics of an online user community,2023 Dominic Detzen