

סמינריון בכריית מידע

"ניתוח נתונים וביצועים של שחקני כדורגל"

מרצה: ד''ר עבדאללה לואי

מגיש:

שמעון זאכי, ת.ז 315767871

**מבוא:**

כדורגל, המשחק הפופולרי ביותר בעולם, אינו רק מקור להנאה ותחרותיות עבור אוהדים ושחקנים כאחד, אלא גם ענף עסקי עצום המוערך בלמעלה מ30 מיליארד יורו. בעולם שבו מועדוני כדורגל משקיעים כ-3 מיליארד יורו מדי שנה ברכישת שחקנים, השאיפה לניצחון בטורנירים ובליגות מהווה דרך לא רק לשפר את תוצאות המשחקים ורזומה המועדון, אלא גם לשפר את המאזן הכלכלי ולמקסם הכנסות מתחרויות שונות, ליגות, חסויות, זכויות שידור ועוד.

מועדונים משקיעים סכומי כסף רבים על מנת לזכות ולנצח במפעלים השונים בכד לגרוף את מירב ההכנסות האפשריות – פרסים, זכויות שידור חסויות ועוד. עם הרצון העז לנצח ולזכות בתארים, כענף ספורט הדברים עלולים ללכת בכיוון הלא רצוי שהוא הפסדים ואי זכיות בתחרויות. כתוצאה מכך, הפסדים אלו יכולים להשפיע באופן ישיר על היכולת לייצר רווחים ולגייס הכנסות נוספות. קבוצות רבות משקיעות אחוזים רבים מסך תקציבם השנתי לרכישת ואימון שחקנים בכדי לעזור לקבוצה לנצח. אך עם זאת, רכישות של שחקנים שאינם מתאימים באופן מדויק לצרכים הספורטיביים של הקבוצה עלולות להוות בזבוז תקציבי נכבד. יתרה מזו, ירידה ברמה הספורטיבית, שתוביל לירידה בליגה, עלולה לגרום לאיבוד חסויות והכנסות מזכויות שידור. לדוגמא, בחינת מסמך מטעם ארגון אופ"א, ארגון הכדורגל הרשמי של התאחדויות הכדורגל האירופאיות, מלמד כי ההכנסות הכוללות הזמינות לחלוקה למועדונים בעונת 2022/2023 הן 2.732 מיליארד יורו בתחרויות השונות, מתוכם 2.032 מיליארד יורו מיועדים למועדונים המתחרים בליגת האלופות ובסופר קאפ של אופ"א שהן התחרויות היוקרתיות והנחשקות ביותר למועדוני ספורט. הצלחה ספורטיבית משפיעה ישירות על המיקום במדרגות הפרסים וההכנסות, כאשר ככל שההצלחה גדולה יותר, כך גם ההכנסות עולות.

כיום, שחקני כדורגל מייצרים נתונים בהיקפים עצומים בכל משחק. כל שחקן מצויד ברצועת זרוע עם חיישן מיקום המתעד את כל תנועותיו במגרש, מהירות, ספרינטים, ומפות חום של מיקומיו השכיחים ביותר. בנוסף, כל פעילות במגרש, בין אם עם הכדור או בלעדיו, מתועדת ונחקרת בכדי להפיק ניתוחים מתקדמים של הביצועים השונים של השחקן והקבוצה. נתונים שנאספים כגון מסירות, נגיעות בכדור, חטיפות ובעיטות, כל אלה משמשים כבסיס נתוננים לניתוחים מתקדמים ביותר שחלקם מבוססים בינה מלאכותית, מה שמאפשר להסיק מסקנות חשובות ולשפר את היכולת הקבוצתית והאישית של כל שחקן.

**הבעיות העסקיות:**

בענף הכדורגל, שהינו ענף תחרותי גם בהיבט העסקי וגם בהיבט הספורטיבי, נתקלים מועדוני הכדורגל באתגרים עסקיים מרכזיים שעלולים להשפיע על היציבות הכלכלית שלהם. כחלק מאסטרטגיית העסק, מועדונים אלו נדרשים להשקיע סכומים נכבדים ברכישת שחקנים, תוך שמירה על יעילות וניצול מיטבי של התקציב הזמין. הצלחה או כישלון בפן הספורטיבי יכולים להשפיע ישירות על זרמי ההכנסות המרכזיים הראשיים של המועדון, כולל חסויות, זכויות שידור והשתתפות בתחרויות בינלאומיות. שחקני כדורגל היום מתפקדים כנכסים עסקיים לכל דבר והינם בעלי שווי שוק המתעדכן באופן שוטף. במסגרת זו, רכישת שחקן כדורגל אינה שונה מכל השקעה עסקית אחרת: היא דורשת תכנון פיננסי מוקפד, אנליזה של תחזיות עתידיות לביצועים, והערכת הסיכונים הכרוכים. שחקנים אלו, בעלי עלויות רכישה ומשכורות גבוהות, מהווים השקעה כלכלית משמעותית, והתמורה עליה.

כיום, המציאות של מועדוני הכדורגל בה הם נדרשים להתמודד עם שוק וענף תחרותי במיוחד, שבו השאיפה להצלחה מתממשת לא רק ביכולת הספורטיבית אלא גם בחוכמת הניהול העסקי והאסטרטגי. המועדונים מחפשים אחר שיטות חדשניות לשפר את הביצועים, להבטיח השגת תוצאות מרשימות ולהבטיח הישגים ארוכי טווח. השימוש בניתוח נתונים ובכלים של בינה מלאכותית נהיה יותר ויותר נפוץ בניסיון להשיג יתרון תחרותי. קבוצות שונות מחזיקות אנשי אנליזה ולעיתים אף מחלקות שלמות שכל מטרתם לנתח ביצועים של קבוצתם, קבוצות יריבות, שחקנים פוטנציאלים ואקדמיית הנוער של המועדון. יתרה מכך, ענקיות טכנולוגיות הצטרפו למהפכה ולמשל חברת Microsoft, ענקית הטכנולוגיה ובעלת דריסת רגל עצומה בתחום הבינה מלאכותית, משמשת כנותנת חסות בליגה הספרדית ומספקת ניתוחים בלייב לצופים בבית ומשדרת את המסקנות על גבי המסך.

המטרות העסקיות של מועדוני הכדורגל הן רבות ומגוונות, אך בסופו של דבר כולן מכוונות לפיתוח הקבוצה ולהגדלת ההכנסות. שיפור ההכנה למשחקים וביצועים אופטימליים במגרש יכולים להביא להישגים ספורטיביים גבוהים יותר, לזכייה בטורנירים שונים והכי חשוב המשכיות ברמה המקצועית והתחרותית בכדי לשמר את גובה ההכנסות. מה שיגדיל בסופו של דבר את זרמי ההכנסות של המועדון מזכויות שידור, חסויות, כרטיסים ועוד. כמו כן, השקעה נבונה ומושכלת ברכישת שחקנים, המבוססת על ניתוח נתונים מדויק ומתקדם, תוכל למנוע בזבוזי תקציב ולהבטיח תמורה מרבית עבור הכסף שהושקע. ישנן מועדונים שמשתמשים בפתרונות מבוססים בינה מלאכותית בכדי לאתר כשרונות צעירים, לפתח את יכולתם המקצועית ולאחר מכך למכור אותם ברווח משמעותי ביחס להשקעה.

**החיבור לבינה מלאכותית וכריית מידע:**

הפתרון של בינה מלאכותית מאפשר למועדוני כדורגל לנתח ולהבין בצורה עמוקה יותר את הנתונים העצומים שנאספים במהלך המשחקים והאימונים. באמצעות שימוש במודלים מתקדמים של למידת מכונה וניתוח נתונים, מועדונים יכולים לקבל תובנות על היכולות האמיתיות של השחקנים, להתאים את האסטרטגיות בהתאם לסגנון המשחק של היריבות, וגלות חולשות ספורטיביות שעדיין לא נוצלו במלואם. בנוסף, ניתוחים אלו יכולים לסייע בבניית תכניות אימון ממוקדות יותר ובהכנה מדויקת למשחקים קריטיים.

העולם המודרני של הספורט, ובפרט הכדורגל, מציב בפנינו אתגרים חדשים בניתוח המספרים וכמות הנתונים עצומה הנערמת בכל משחק. שחקני כדורגל, בזכות ניתוח וידאו וטכנולוגיות מתקדמות הכוללות חיישנים חכמים המשתלבים באספקט המשחק ומתעדים את מיקומי השחקנים ופעוליתיהם, מפיקים כמויות נתונים עצומות המתעדות ומייצרות כמויות נתונים בנפחים עצומים. כל פעולה עם הכדור או בלעדיו נתונה לניתוח, מה שמעניק לנו הזדמנות להשתמש בכלים מבוססים בינה מלאכותית כדי להסיק מסקנות מעמיקות ולשפר את הביצועים הקבוצתיים והאישיים.

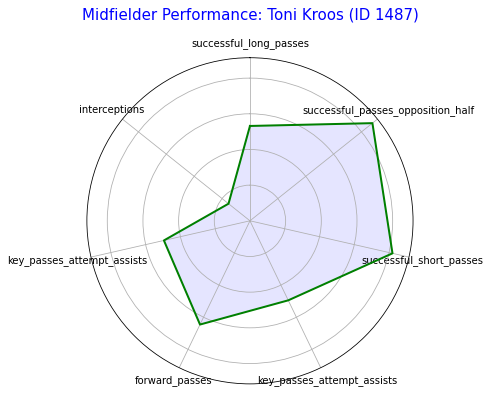
המטרה העיקרית של כריית הנתונים והשימוש בבינה מלאכותית בעולם הכדורגל היא לחדור אל מעמקי המידע הזמין ולהשיג תובנות שיכולות להפוך ליתרון תחרותי ברור. זאת נעשית באמצעות מודלים מתקדמים המאפשרים לחזות תוצאות, לנתח את התרומה של כל שחקן לקבוצה, להתאים אסטרטגיות משחק ולפתח תוכניות אימון מדויקות. הפלט מכל תהליך זה הוא מערך של ניתוחים, תחזיות והמלצות שמתבססות על נתונים אמינים ומדויקים, מה שמאפשר למועדונים לקבל החלטות מושכלות יותר בכל הקשור לאסטרטגיה הספורטיבית והעסקית.

בכדי להשיג את היעדים הללו, נעשה שימוש במגוון טכנולוגיות ומודלים של בינה מלאכותית. השילוב של מידע כמותי עם טכניקות אנליטיות מתקדמות מאפשר לנו להבין טוב יותר את המשחק ולשפר את הביצועים של השחקנים והקבוצה כולה. ניתן לבצע זאת על ידי מספר כלים שניתן לפתח.

* **מודל תחזית:** לתכנן וליצור בעזרת מודלים סטיסטים ולמידת מכונה,
* **פרופיל שחקנים**: לבנות פרופיל מותאם אישית עבור השחקנים על ידי בחירת המאפיינים הרצויים, יסייע לנו להבין את סט הכישורים והיכולות של השחקן.
* **התאמת תפקידים**: היום, סט הכישורים של השחקנים רחב יותר. השחקנים יכולים להתאים למספר תפקידים, שחקן הגנה יכול לשחק כקשר אמצע ולהביא את איכות ההגנה שלו לתפקיד של קשר אמצע.

לדוגמא, בניית פרופיל אישי לכל שחקן בהתאם לתפקידו, יכולה לעזור לנו לאפיין את סט היכולות של השחקן.

תמונה שמכילה תרשים, קו, עיגול, אוריגמי

התיאור נוצר באופן אוטומטי

בשני הרדארים שמוצגים, אנו יכולים לראות שני קשרים באותה קבוצה בעלי סגנון שונה. אחד מהם עושה יותר פעולות הגנתיות על מנת לחטוף את הכדור מהיריבה ולהעביר אותו קדימה (ואלוורדה) והשני הוא קשר קלאסי אשר מטרתו במשחק היא לנווט את הכדור אל השחקנים השונים ולנהל את המשחק(קרוס).

**הנתונים:**

עבודה זו מבוססת על מערך נתונים של שחקני כדורגל משתי הליגות הבכירות בספרד, "לה ליגה" ו"לה ליגה 2", במהלך עונת המשחקים של 2021-2022. הנתונים אוספו דרך תצפיות ומדידות בכל אחד מ-38 המשחקים של העונה הרגילה. הדגימות כללו מגוון רחב של למעלה מ130 פיצ'רים, החל מפרטים ביוגרפיים וכלה בנתונים סטטיסטיים מתקדמים אודות ביצועיהם במגרש. לצורך העבודה, החלטתי להשמיט את דגימות ליגת הנשים מכיוון שההבדלים בין נתוני גברים ונשים דורשים ניתוח נפרד ומסוימת. סט הנתונים נדגם ונלקח מפלטפורמת Kaggle. קישור לסט הנתונים מופיע ברישום הביבליוגרפי.

איכות הנתונים ודיוק המידע הם הבסיס לכל ניתוח סטטיסטי או מודל בינה מלאכותית. נתונים שאינם מדויקים או אינם מייצגים את המציאות באופן נאות עלולים להוביל לחיזויים שגויים ולהחלטות מבוססות על נתונים מוטעים. לכן, בפרויקט זה מושם דגש עצום על בדיקה קפדנית של הנתונים לפני ולאחר אסיפתם. האתגר העיקרי הוא לבנות סט נתונים מאוזן ומקיף, שיכלול את כל המשתנים הרלוונטיים וישקף את הפרמטרים הקריטיים לניתוח. אימות הנתונים בוצע אל מול האתר הרשמי של הליגה הספרדית ואתרי ספורט שונים ברחבי הרשת האוספים.

הפעולות שבוצעו על סט הנתונים כוללות:

השלמת ערכים חסרים: כחלק מחלוקת התפקידים על המשחק, נוצר מצב שבו חלק מהפיצ'רים נשארים ריקים והם הושלמו בערך 0.

נירמול סט הנתונים: בכדי לקבל תוצאות מדוייקות ולא מוטות מפיצ'רים כמותיות שונים בעלי ערך גבוה, כל סט הנתונים מנורמל באמצעות נירמול מינימום מקסימום.

בנוסף, סט הנתונים פוצל לשניים, סט נתוני פעילות הכולל את הערכים הכמותיים של כל פיצ'ר ואת מספר התצפית ולסט נוסף המכיל את מספר התצפית ופרטים אישיים על התצפית כמו שם מלא, קבוצה בליגה וכו'.

**מודלים ומימושם:**

המודלים שנבחנו לטיפול בבעיות שעלו:

**Ridge (L2):** רגרסיה רידג' היא גישה ברגרסיה לינארית שבה יש תוספת של עונש על מקדמי המשתנים במודל. תוספת זו נועדה להפחית את הסיכון להתאמה יתר שעלולה להתרחש כאשר מקדמים מסוימים גדולים מדי. כתוצאה מכך, המודל נוטה להיות יותר יציב ולבחור במשתנים הרלוונטיים ביותר (הסטי ט', טיבשיראני ר', ופידמן ג', 2008).

**Lasso (L1):** דגם רגרסיה זה מכניס עונש על מקדמי המשתנים בדומה לרידג', אך עשוי להוביל לאפס מקדמים למשתנים פחות חשובים. זה מאפשר בחירה אוטומטית של משתנים ומסייע למנוע התאמה יתר במודל (הסטי ט', טיבשיראני ר', ופידמן ג', 2008).

**ElasticNet:** מודל זה משלב בין רגרסיה רידג' ולאסו, נועד להציע את היתרונות של שניהם. הוא מאפשר למודל לבחור את המשתנים הרלוונטיים ולהציג יציבות גם במקרים של נתונים עם קורלציות גבוהות או רב ממדיות (הסטי ט', טיבשיראני ר', ופידמן ג', 2008).

**RandomForestRegressor**: רגרסיית יער אקראי הינו מודל של יער אקראי שמשלב מספר רב של עצי החלטה כדי להגיע לחיזוי מדויק ויציב ככל הניתן. המודל בונה כל עץ החלטה באופן עצמאי ואקראי, ולבסוף משלב את תוצאות החיזוי של כל העצים לחיזוי הסופי.

**XGBooster:** אלגוריתם שמטרתו לייצר סדרה של דגמים (בדרך כלל עצי החלטה) שמשפרים זה את זה באופן איטרטיבי. כל דגם חדש מנסה לתקן את השגיאות של הדגם הקודם, ובכך לשפר את החיזויים בכל פעם.

**האלגוריתמים שנבחרו:**

**Ridge Regression**: רגרסיה רידג' (Ridge Regression): זוהי שיטת רגרסיה לינארית המתמודדת עם בעיית הריבוי גורמים בנתונים. בשיטה זו, מוסיפים למודל "עונש" על מקדמים גבוהים מדי, מה שעוזר לשמור עליהם בגודל סביר ולמנוע מצב שבו המודל מתאים יתר על המידה לנתוני האימון. השיטה מאפשרת לזהות את המשתנים החשובים ביותר עבור תחזית מסוימת, תוך התעלמות ממשתנים שאינם מוסיפים מידע רלוונטי. השתמשתי ברגרסיית רידג' לבחרית פיצ'רים אלו, בכדי לחזות באופן רציף תוך שימוש ברגולריזציה כדי למנוע אוברפיטינג. השימוש ברגרסיית רידג מאפשר לאתר את הפיצ'רים המשפיעים ביותר על מספר השערים ששחקן כובש תוך התעלמות מפיצ'רים שאינם רלוונטים או ישירות קשורים לשערים.

**K-Fold Cross-Validation:** שיטה זו מחלקת את הנתונים ל-K חלקים שווים, כשבכל פעם אחת מהקבוצות משמשת כקבוצת בדיקה והשאר כקבוצת אימון. כל נתון משתתף בדיוק פעם אחת בקבוצת הבדיקה ו-K-1 פעמים בקבוצת האימון. השיטה מאפשרת הערכה יותר מדויקת ואובייקטיבית של ביצועי המודל, משום שהיא נותנת לו להתאמן ולהיבדק על כל הנתונים. בחירה ב-K גבוה מאפשרת רזולוציה גבוהה יותר של ההערכה, אך דורשת יותר משאבי חישוב (ברק ר' 2017). בחרתי להשתמש באלגוריתם הזה על מנת לשפר את יכול האימון של המודלים ולהימנע מאוברפיטינג מכוון שאנו משנים את סט נתוני האימון בכל פעם. בחרתי ב **K=12** מכוון שקיבלתי איתו את ערכי R², MSE, RMSE הטובים ביותר לשאר ערכי K שנבדקו.

Random Forest Regressor: הוא אלגוריתם שמשתמש במספר רב של עצי החלטה בכדי להוציא תחזית. גרסתו לבעיות רגרסיה, Random Forest Regressor, פועלת על ידי אימון מספר עצי החלטה על תת-קבוצות של הנתונים (בדרך כלל באופן אקראי ועם חזרות) ואז מתבצעת הערכה ממוצעת של התחזית מכל העצים כדי לקבל את התחזית הסופית. זה מקטין את הסיכון לאובר-פיטינג מכיוון שהוא מבצע רגולריזציה באופן טבעי על ידי אינטגרציה של תוצאות ממגוון עצים  
(מטלוף נ' 2017). מספר העצים שנבחרו בהרצת מודל זה הינו **N=250** במספר הזה האלגוריתם מתכנס ומתקבלות תוצאות המדדים הטובות ביותר. בנוסף, זהו המודל שנבחר כיעיל ואפקטיבי ביותר משאר המודלים עם תוצאות המדדים הטובות ביותר.

פלט דוגמא:

**Random 10 Unique Predictions with More Than 5 Predicted Goals:  
Player: Alexander Isak, Predicted Goals: 11.573988574482065, Actual Goals: 6  
Player: Ante Budimir, Predicted Goals: 8.128449867270778, Actual Goals: 8  
Player: Lucas Ocampos, Predicted Goals: 5.362365481178296, Actual Goals: 6  
Player: Mohammed Dauda, Predicted Goals: 8.304394478129495, Actual Goals: 9  
Player: Raúl de Tomás, Predicted Goals: 15.177771874870636, Actual Goals: 17  
Player: Bárbara Latorre, Predicted Goals: 7.096, Actual Goals: 7  
Player: Jessica Martínez, Predicted Goals: 5.818811893463135, Actual Goals: 6  
Player: Manuel Barreiro Bustelo, Predicted Goals: 5.459696292877197, Actual Goals: 7  
Player: Anthony Lozano, Predicted Goals: 6.497333156222379, Actual Goals: 7**

**Player: Maximiliano Gómez, Predicted Goals: 6.057991981506348, Actual Goals: 5**

**Player: Karim Benzema, Predicted Goals: 21.00915619072694, Actual Goals: 27**

בעזרת אלגוריתם RandomForestRegressor, נוצרו תחזיות עבור כל שחקן הנעשות על ידי חיזוי הממוצע ממספר עצי החלטה שמופעלים באופן אקראי, כך שכל עץ מחזה את מספר השערים שהשחקן יבקיע. התוצאות המקבלות מבוססות על נתונים היסטוריים ומשקפות את הפוטנציאל הגלום בכל שחקן.

הפלט של המודל מציג תחזיות למספר השערים שכל שחקן צפוי לכבוש, עם חיזויים שנעים בין הדייקניים לאלה שסוטים מעט מהמציאות, המשקפים את האתגרים העיקריים בחיזוי ביצועי ספורטאים. למערכת זו שני סוגי פלט עיקריים:

1. תחזיות כמותיות: כוללות חיזויים עבור כל שחקן על פי נתוני ההיסטוריה שלו, ומתייחסות לפיצ'רים שנמצאו כרלוונטיים לחיזוי הביצועים.

2. אנליזה כוללת: כוללת ניתוח הפערים בין תוצאות החיזוי לביצועים האמיתיים, מה שמסייע לזהות גורמים המשפיעים על הביצועים ולהצביע על אפשרויות לשיפור.

הפלט כולל גם דוחות חיזוי המציגים את התחזיות לכל שחקן עם פרטים על השערים הצפויים, ואנליזה סטטיסטית הכוללת את דיוק החיזויים, שגיאות טיפוסיות ופערים בולטים .בנוסף, המערכת מספקת תובנות עמוקות למאמנים ומנהלי קבוצות, ומאפשרת להם לתכנן אסטרטגיות לעונות הקרובות על בסיס היכולות הפוטנציאליות והקיימות של השחקנים. דוגמה לתוצאות הפלט של המודל ניתן לראות 10 שחקנים שכבשו מעל לחמישה שערים במהלך העונה, ואת מספר השערים שהמודל חזה שיכבשו.

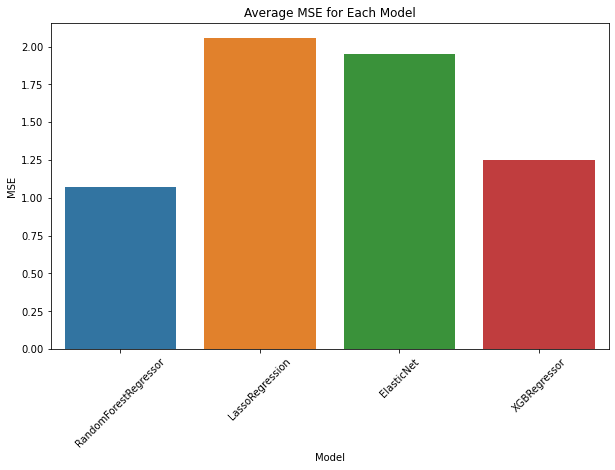
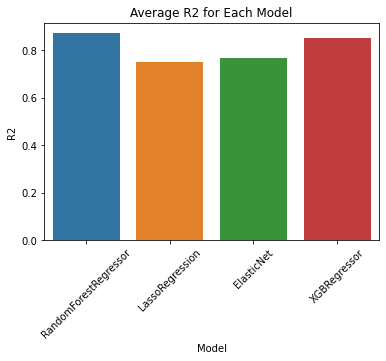
**אנליזה והערכת מודלים:**

לצד תוצאות טובות והפרשים של פחות משער לחלק מהתוצאות, ישנן גם תוצאות פחות טובות כמו Alexander Isak, ו Karim Benzema. השחקנים הללו קיבלו ערכים חזויים שרחוקים מהערך האמיתי מכוון שמדובר בתצפיות חריגות סטטיסטית. למשל, Karim הינו שחקן שמנצל את ההזדמנויות שלו וכובש שערים במס' נמוך של בעיטות למסגרת למשל, שהינו פ'יצר משמעותי במודל. לעומת זאת Isak מנגד, הינו שחקן שמחמיץ המון ויזדקק להמון הזדמנויות בשביל לכבוש שער. זהו תמרור אזהרה לקבוצה שמעוניינת לרכוש אותו כי היא יכולה לרכוש חלוץ שאינו מבקיע גולים ביחס להזדמנויות שחבריו לקבוצה יוצרים לו.

תמונה שמכילה טקסט, צילום מסך, תרשים, עלילה

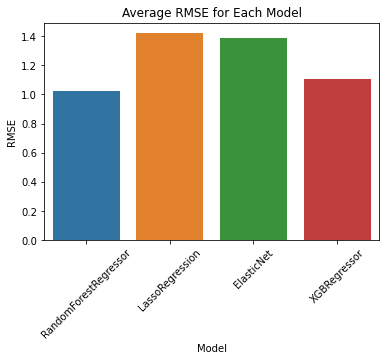
התיאור נוצר באופן אוטומטילפי התוצאות, ניתן לראות שמודל רגרסיית היער האקראי (RandomForestRegressor) הינו המדויק והיעיל ביותר לחיזוי הבקעת שערים לשחקן. זה מתבטא בערכי R², MSE ו-RMSE הגבוהים והנמוכים בהתאמה שלו לעומת שאר המודלים שנבחנו. הערך הממוצע של R², שהוא מדד לכמה טוב המודל מסביר את השונות בנתונים, הוא הגבוה ביותר למודל זה, מה שמראה על דיוק חזותי טוב.

מתוך המודלים האחרים שנבחנו, נראה כי רגרסיית Lasso ו-ElasticNet סבלו מביצועים נמוכים יחסית עם ערכי R² נמוכים יותר, המראים על דיוק חזותי נמוך יותר. ה-XGBRegressor הציג תוצאות טובות עם ערכי R² קרובים לאלו של רגרסיית היער האקראי, אך עדיין קצת נמוכים יותר.

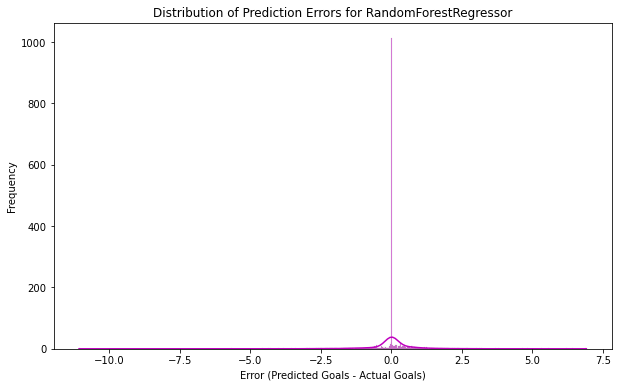


תמונה שמכילה טקסט, צילום מסך, תרשים, עלילה

התיאור נוצר באופן אוטומטי



אפשר להגיד שהמודל מאוזן במידה מסויימת, עם רוב השגיאות מתקבצות סביב הערך 0 בהיסטוגרמה, אך ישנם גם זנבות שמצביעים על קיומן של תחזיות לא מדויקות. אם ההתפלגות הייתה מוטה יותר לכיוון מסוים, זה היה מרמז על חיזוי יתר או מעט לטובת התחזיות.



גרף פיזור הערכים החזויים אל מול הערכים האמיתים.

תמונה שמכילה צילום מסך, טקסט, תרשים

התיאור נוצר באופן אוטומטי

עם זאת, חשוב להתייחס באופן ביקורתי למגבלות הטמונות בכל אחד מהם. לדוגמה, מודלים כמו רגרסיה לינארית עלולים להתקשות בהתמודדות עם המורכבות והדינמיות של נתוני כדורגל, המאופיינים ביחסים לא לינאריים ותלות מורכבת בין משתנים. בנוסף, מודלים המבוססים על עצי החלטה ויערות אקראיים עלולים להיתקל באתגר של יתר התאמה (Overfitting), במיוחד כאשר הם מתמודדים עם נתונים המכילים רעש רב או פיצ'רים לא רלוונטיים. חשוב גם להביא בחשבון את הטיות שעלולות להתקיים בתהליך איסוף הנתונים, כגון נטייה להעדיף שחקנים מליגות מסוימות או עמדות מסוימות, שיכולות להשפיע על תוצאות המודלים ולצמצם את יכולתם לספק תובנות אובייקטיביות.

**סיכום:**

לסיכום, רגרסיית היער האקראי נמצאת כהכי מדויקת לחיזוי הבקעת שערים לפי הנתונים שהוצגו, ומבטיחה חיזויים אמינים ומדויקים ככל הניתן לנתונים הקיימים. שיפור דיוק המודל: ניתן לשפר את דיוק המודל באמצעות הוספת ערכים וסט נתונים נוספים העוסקים בפיצ'רים נוספים שלא נמצאים באוסף הנתונים שעליו העבודה מתבססת. נתונים נוספים יכולים להיות מפות חום שכיחות של שחקנים, תאוצות והתנהגויות נוספות עם וללא הכדור יכולות לתרום רבות לדיוק המודל.  
את המודל ניתן להטמיע בקלות בארגונים ובמועדונים באמצעות דוחות שמתקבלים על ידי התחרויות השונות על ביצועי השחקנים, שימוש באמצעים טכנולוגים כמו חיישני כדור ורצועות חיישני מיקום ושימוש בממשק תכנות יישומים לכריית מידע מאתרי תוצאות ונתוני ספורט שוטפים או חוזי שותפות איתם בהתאם להגבלות המשפטיות. כאשר מתבוננים בהשלכות הרחבות של המחקר על מועדוני כדורגל, מאמנים ותעשיית הספורט בכללותה. כיצד ניתן להשתמש בממצאים כדי לשפר אסטרטגיות אימון, לעצב תוכניות אימונים מותאמות אישית עבור שחקנים, או להקנות לצוות המאמנים כלים להכרה וטיפוח כשרונות? בנוסף, יישום ניתוח מתקדם זה עשוי להציע יתרון תחרותי בתהליכי סקאוטינג ורכש של שחקנים, על ידי הזהות שחקנים עם פוטנציאל גלום המתבטא בנתונים ולאו דווקא בביצועים הנראים לעין. העברת תובנות אלו לשטח, והתאמתן לאסטרטגיה ולמטרות הספציפיות של כל מועדון, יכולה להגביר את היעילות ואת ההצלחה של קבוצות כדורגל ברחבי העולם.

ניתן לראות כי אפשר וכדאי להשתמש בכלים המבוססים בינה מלאכותית בכדי לנתח, לחקור ולשפר את הביצועים של שחקני כדורגל. עם זאת, מומלץ לאסוף נתונים נוספים, כגון מיקומי השחקנים לאורך המשחק, פעולות נוספות ללא הכדור ונתונים קבוצתיים בכדי לבצע הערכות נוספות ואפילו מדויקות יותר גם בענף ספורט בעל המון משתנים תלויים ובלתי תלויים כמו כדורגל.  
בסך הכל אני מרוצה מדיוק המודל אף על פי שניתן לשפר. נהניתי מעשיית העבודה ולמדתי המון על המודלים והכלים השונים בהם התבססתי בעבודה.

**ביבליוגרפיה:**

1. סט הנתונים – La-Liga Stats 2021-2022 קישור [כאן](https://www.kaggle.com/datasets/sdelquin/laliga-data)
2. קישור למסמך מטעם אופ''א – קישור [כאן](https://editorial.uefa.com/resources/0277-158b0bea495a-ba6c18158cd3-1000/20220704_circular_2022_47_en.pdf)
3. Hastie T, Tibshirani R, Friedman J , Elements of Statistical Learning, 2nd Edition, Springer, 2008
4. Berk R, Statistical Learning from a Regression Perspective, 2nd Edition, Springer, 2017
5. [3] Matloff N, Statistical Regression and Classification: From Linear Models to Machine Learning