השפעת רמת השליטה בכדור במחצית הראשונה על יכולת הקבוצה במחצית השנייה

מגישים:

רועי רימר 314828732

משה עבאדי 324658939

קישור ל-repository:

https://github.com/moshinhoabadi/Causal-Inference-Project

מבוא - הצגת הבעיה

בכדורגל, ישנו מגוון רחב של אסטרטגיות שקבוצות נוקטות על מנת לנסות ולנצח את יריבותיהן. יחס האסטרטגיות השונות בנוגע להחזקה בכדור הוא מגוון גם כן. על קצה אחד של הספקטרום, ישנן שיטות הדוגלות בהחזקה בכדור זמן רב ככל הניתן, כך שהקבוצה תוכל ליזום ולקבוע את קצב המשחק. מן העבר השני, ישנן אסטרטגיות בהן הקבוצה מאפשרת ליריב להחזיק בכדור וליזום בעצמו, אך "אורבת" לכל טעות שלו ולכל איבוד כדור פזיז מתוך כוונה לצאת במקרה זה להתקפה מתפרצת ולתפוס את הגנת היריב לא מוכנה.
השליטה בכדור עשויה להיות תלויה גם ביכולת הטכנית והארגונית של הקבוצה ושל שחקניה להתמסר ביניהם ולהחזיק בכדור מבלי להיכנע לניסיונות החטיפה והלחץ של היריבה ולאבד אותו.

הקבוצות עלולות לנהוג באופן דינאמי במהלך המשחק, ולשנות את הקצב בו הן משחקות או את מידת הניסיון שלהן להחזיק בכדור לפי שיטות מסוימות או לפי הדרך בה מתנהל המשחק. לדוגמא, ייתכן כי קבוצה ששלטה בכדור במחצית הראשונה, תתעייף ותהיה פאסיבית יותר במחצית השנייה, או שמא תתגונן בכוונה בכדי לא לתת לניצחון לחמוק במידה והיא מובילה. מנגד, קבוצה אחרת יכולה לנסות במקרה כזה דווקא להמשיך ולנסות להחזיק בכדור כדי "לא להוריד את הרגל מהגז" ולאפשר ליריבתה לחזור לעניינים.

בעבודה שלנו, בחרנו לנסות ולמדוד את האפקט הקוזאלי של השליטה בכדור במחצית הראשונה של המשחק, על יכולת הקבוצות במחצית השנייה, אותה בחרנו למדוד באמצעות כמות הבעיטות למסגרת שבעטה הקבוצה במהלך המחצית השנייה, במדידת האפקט הקוזאלי אנו רוצים לנסות ולמצוא את רמת ההשפעה שיש לאופן בו התפתחה המחצית הראשונה על המחצית השנייה, ולראות כיצד הקבוצות מגיבות למהלך המשחק. האם הקבוצה שהחזיקה בכדור יותר במהלך המחצית הראשונה תהיה בעלת מחץ התקפי במחצית השנייה? האם הקבוצה שהייתה פאסיבית יותר במחצית הראשונה תנסה במהלך המחצית השנייה להתקיף וליזום יותר?

הנתונים

הנתונים בהם אנו משתמשים לקוחים מתוך ה-European Soccer Database[1] ומתוך סט הנתונים המשלים European Soccer Database Supplementary[2] הנמצאים באתר Kaggle.
הנתונים בהם אנחנו משתמשים מכילים מידע על כ-5500 משחקי כדורגל שנערכו בין השנים 2008-2016 במסגרת שש ליגות כדורגל בכירות באירופה (ליגות הכדורגל הבכירות של אנגליה, ספרד, גרמניה, איטליה, צרפת והולנד). מסד הנתונים מורכב מכמה קבצים המתייחסים למידע על משחקים, קבוצות או שחקנים.

המידע עבור כל משחק מכיל את הנתונים הבאים:

- זהות הקבוצות שנטלו חלק במשחק.
- ההרכבים הפותחים של שתי הקבוצות.
- הליגה בה התקיים המשחק.
- התאריך, העונה ומספר המחזור בהם התקיים המשחק.
- מאורעות שקרו במהלך המשחק: בעיטות למסגרת, שערים, עבירות, כרטיסים ואחוזי ההחזקה בכדור ברגעים מסוימים במשחק.
- יחסי הימורים עבור ניצחון של קבוצת הבית, של קבוצת החוץ ועבור תיקו מתוך תשע סוכנויות הימורים שונות.

המידע עבור כל קבוצה מכיל את הנתונים הבאים:

- שם הקבוצה.
- נתונים עבור סגנון המשחק של הקבוצה [3], כמו למשל רמת האגרסיביות בהגנה, המהירות בה הקבוצה נוטה לבנות התקפות, אופי סוגי המסירות שהקבוצה נוהגת למסור ועוד. נתונים אלו לקוחים מתוך משחק המחשב FIFA, ומעודכנים עבור כל קבוצה לכמה נקודת זמן שונות בין השנים המופיעות בסט הנתונים.

המידע עבור כל שחקן מכיל את הנתונים הבאים:

- שם השחקן.
- נתונים פיזיים כגון גיל, גובה ומשקל.
- מדדים שונים של איכות השחקן, כמו למשל יכולת הבעיטה שלו, רמת השליטה בכדור, והדירוג הכללי (overall rating) של השחקן, הלוקח בחשבון את כל המדדים האחרים ומביא אותם לכדי ציון משוקלל המתחשב גם בעמדה הטקטית טבעית של השחקן. מדדים אלו לקוחים גם הם מתוך משחק המחשב FIFA, ומעודכנים עבור כל שחקן לכמה נקודות זמן שונות בין השנים המופיעות בסט הנתונים.

נציין שני אתגרים הקשורים לניתוח סט הנתונים הנ"ל:

1. הנתונים והמדדים השונים שיש עבור השחקנים והקבוצות מעודכנים לכמה נקודות זמן שונות, אך הם אינם משויכים מראש לכל משחק, אלא נמצאים בקובץ נפרד. יש לאתר לכל משחק את עדכון הנתונים המתאים ביותר של הקבוצות והשחקנים המשתתפים, ולשייך אותם עבור אותו המשחק.
2. מסד הנתונים בו אנו משתמשים מכיל קבצים רבים, וכל אחד מחזיק חלק אחר מהנתונים. למשל, יש קובץ אחד המכיל נתונים מסוימים על המשחקים, קובץ אחר המכיל נתונים על הקבוצות, קובץ אחר המכיל נתונים על אירועי בעיטות למסגרת שקרו במשחקים השונים ועוד. אנו נבצע אינטגרציה של הנתונים הללו לכדי סט נתונים אחד שיכיל עבור כל משחק רלוונטי את כלל ה-covariates בהם נשתמש לצורך המחקר.

תכנון השאלה הקוזאלית והמחקר

השאלה הקוזאלית אותה אנו רוצים לחקור היא: **מהי ההשפעה של רמת השליטה בכדור במחצית הראשונה של המשחק, על יכולת הקבוצה במחצית השנייה?**
את היכולת בחרנו להגדיר ככמות הבעיטות למסגרת שבעטה הקבוצה במחצית השנייה, והחלוקה לקבוצות המחקר נעשתה לפי אחוזי השליטה בכדור במחצית הראשונה. אנו ננסה למדוד את ה-ATE, כלומר את ה-Average Treatment Effect.

הסיבה לבחירה בבעיטות במסגרת כמדד ליכולת היא שבעיטות אלו הן בעיטות שהיו נכנסות לשער אילולא היו נבלמות על ידי השוער או שחקן הגנה אחר. לכן, בעיטות למסגרת נוטות להיות מסוכנות יותר מאשר בעיטות שהן לא בעיטות למסגרת. בחרנו למדוד בעיטות למסגרת ולא שערים על מנת לנטרל את ההשפעה של מקרים בהם למשל הקבוצה הייתה מסוכנת ובעטה הרבה למסגרת, אך השוער בקבוצה השנייה היה מעולה ועצר את הבעיטות. במקרים כאלו נרצה עדיין להתחשב בכך שהקבוצה הייתה מסוכנת, אף על פי שהדבר לא בא לידי ביטוי בהכרח בכמות השערים שהיא כבשה בפועל.

כדי להגדיר בצורה טובה יותר את המחקר שנבצע לצורך מענה על שאלת המחקר, נשתמש קודם כל בפרדיגמת ה-target trial, וננסה לתאר איך היינו בונים ניסוי RCT שינסה לענות על השאלה הקוזאלית שהגדרנו.

עבור ה-Target Trial:

Eligibility criteria: אילו היינו יכולים לשלוט בכך, היינו בוחרים להכליל בניסוי כמות גבוהה של משחקים על פני כמה שנים, כאשר בכל משחק תוגרל קבוצה לשחק עם קבוצה אחרת הדומה ברמתה מבחינת איכות השחקנים בשתי הקבוצות. נמנע אפשרות שבה תוצאת המשחק חשובה יותר לאחת הקבוצות מאשר לקבוצה השנייה.

Treatment strategies: בכל משחק מבין המשחקים בניסוי, היינו מגרילים עבור כל משחק באופן אקראי זוג אסטרטגיות, אחת לכל קבוצה, בהן ינקטו הקבוצות במהלך המחצית הראשונה. האסטרטגיות יגדירו את היחס של כל אחת מן הקבוצות לגבי רמת ההחזקה בכדור. נגדיר את זוגות האסטרטגיות כך שפער אחוז השליטה בכדור בין שתי הקבוצות במחצית הראשונה יהיה משמעותי (למשל שקבוצה אחת תחזיק בכדור לפחות 60% אחוז מהזמן במחצית הראשונה).

Assignment procedures: כאמור, בכל משחק יוגדר לכל קבוצה באקראי כיצד עליה לשחק במחצית הראשונה מבחינת יחס הקבוצה לרמת ההחזקה בכדור. הקבוצות יהיו מודעות לאסטרטגיה שהן צריכות לשחק לפיה. הקבוצה שיוגדר לה לנסות ולהחזיק בכדור יותר תהיה שייכת לקבוצת מחקר $T = 1$, ואילו הקבוצה השנייה תהיה שייכת לקבוצת המחקר $T = 0$.

Outcome: בכל משחק, נמדוד עבור כל קבוצה את כמות הבעיטות למסגרת שהיא בעטה במהלך המחצית השנייה.

Analysis plan: מכיוון שזהו ניסוי RCT, כדי לקבל את ה-ATE הרצוי נוכל לחסר את ממוצע הבעיטות למסגרת במחצית השנייה עבור קבוצות ששלטו בהחזקה בכדור במחצית הראשונה בממוצע הבעיטות למסגרת במחצית השנייה עבור קבוצות שלא שלטו בהחזקה בכדור במחצית הראשונה, ונקבל את האפקט הקוזאלי הרצוי. עם זאת, עבור כל משחק נגריל באקראי קבוצה אחת בלבד מבין השתיים ונשתמש ב-outcome שהתקבל עבורה, וזאת על מנת למנוע תלות בין חלק מה-outcomes.

כעת, לאחר שהגדרנו את ה-target trial, נשתמש בו על מנת להגדיר כיצד נבצע את ה-observational study שלנו. אנו כמובן איננו יכולים לשלוט באסטרטגיות של הקבוצות, ברמת השליטה שלהן בכדור, או בהפרש הרמות בין הקבוצות המשחקות, אך ננסה להתחשב בדברים האלו בהגדרת המחקר, ולסנן משחקים לא רלוונטיים במידת הצורך.

פרטים עבור ה-Observational Study:

Eligibility criteria: משחקים בין קבוצות בין השנים 2008-2016, כאשר נשתמש רק בנתונים עבור משחקים שלא התקיימו באחד משני המחזורים האחרונים באותה עונה (כדי לנסות לצמצם מקרים של משחקים "לפרוטוקול בלבד"), בהן אחת הקבוצות החזיקה בכדור לפחות 60% מזמן המחצית הראשונה, וכאשר הפרשי הרמות בין הקבוצות שמשחקות אינו גבוה (הסבר על חישוב בהמשך). בנוסף, לא נכלול גם משחקים בהן אחת הקבוצות קיבלה כרטיס אדום, מכיוון שכרטיס אדום עשוי לשנות את מהלך המשחק באופן די דרסטי, כמו גם את הפרשי הרמות בין הקבוצות (שכן אחת הקבוצות משחקת עם שחקן שדה אחת פחות) ועל הצורה בה כל קבוצה תשחק.

Treatment strategies: נגדיר טיפול בינארי באמצעות אינדיקטור T, כך ש-$T = 1$ מסמל שהקבוצה החזיקה בכדור לפחות 60% מזמן המחצית הראשונה, ו-$T = 0$ מסמל שהקבוצה החזיקה בכדור לכל היותר 40% מזמן המחצית הראשונה.

Assignment procedures: בכל משחק מבין המשחקים בניסוי נתייחס לקבוצה שהחזיקה בכדור לפחות 60% מהמחצית הראשונה כשייכת לקבוצה בה T=1, ואילו את הקבוצה השנייה נשייך לקבוצה בה T=0. כאמור נשתמש רק במשחקים בהם הייתה קבוצה שהחזיקה בכדור לפחות 60% מזמן המחצית הראשונה.

Outcome: בכל משחק, נבחן עבור הקבוצה עליה בחרנו להסתכל את כמות הבעיטות למסגרת שהיא בעטה במהלך המחצית השנייה.

Analysis plan: נחשב את ה-ATE הרצוי באמצעות שיטות שונות: S-Learner, T-Learner, Matching, IPW. מכל משחק, נבחר באקראי את אחת הקבוצות ונשתמש רק בנתונים עליה. בדרך זו נמנע כפילות וקשר בין הנתונים, בהם יהיו שתי רשומות דומות מאוד מכיוון שאלו קבוצות ששיחקו זו מול זו באותו משחק.
נבצע את התהליך עבור קומבינציות אקראיות שונות של בחירת קבוצה מכל משחק, ונמצע עבור כל שיטה את ה-ATE שהתקבל על פני כל הקומבינציות.

פירוט - Measured and Unmeasured Confounders

להלן ה-confounders בהם השתמשנו בשיטות השונות לחישוב ה-ATE. Confounders אלו יהיו הפיצ'רים של תצפית ספציפית. כל משחק ייוצג על ידי תצפית אחת, שתהיה מנקודת מבט של אחת משתי הקבוצות שהשתתפו בו.

- האם הקבוצה הייתה קבוצת הבית.
- כמות הבעיטות למסגרת שבעטה כל קבוצה במחצית הראשונה.
- כמות השערים שהבקיעה כל קבוצה במחצית הראשונה.
- הליגה בה התקיים המשחק.
- מספר מחזור הליגה בו התקיים המשחק.
- עונת המשחקים בה התקיים המשחק.
- החודש בשנה בו התקיים המשחק.
- כמות העבירות והכרטיסים הצהובים שהיו לכל קבוצה במחצית הראשונה.
- יחסי ההימורים לניצחון של הקבוצה, יחסי ההימורים לניצחון של היריבה, ויחסי ההימורים לתיקו. היחסים מחושבים על ידי ממוצע של יחסי ההימורים על סמך תשע סוכנויות הימורים שונות.
- נתונים על אופי שיטת המשחק של הקבוצה, המחושבים על סמך המידע הנתון מ-FIFA (מבין אלו שבמסד הנתונים) הקרוב ביותר קלנדרית למועד המשחק עבור אותה הקבוצה. כנ"ל עבור נתוני אופי המשחק של הקבוצה היריבה. נתונים אלו כוללים למשל את הנטייה של כל קבוצה הקבוצה לשחק במסירות ארוכות, רמת הלחץ ההגנתי שהקבוצה נוטה לבצע, הנטייה שלה לנסות ולהחזיק בכדור ועוד.
- רמת שחקני השדה של הקבוצה על הנייר על סמך סגל השחקנים שלה, וכנ"ל עבור הקבוצה היריבה. נתונים אלו חושבו באופן הבא: עבור כל קבוצה וכל עונה, הסתכלנו על הרכבי הקבוצה עבור כל המשחקים באותה עונה. לאחר מכן, עבור כל שחקן לקחנו את נתוני דירוג ה-overall rating מתוך המידע הנתון מ-FIFA (מבין אלו שבמסד הנתונים) הקרוב ביותר קלנדרית עבור אותו שחקן לעונה בה התקיים המשחק. מיצענו את הנתון

הזה על פני כל ההרכבים שהופיעו באותה עונה, וכך קיבלנו מדד המשקף את הרמה הממוצעת של שחקני השדה של הקבוצה, הלוקחת בחשבון את השחקנים השונים בהם ששיחקו בהרכב הקבוצה במהלך העונה ואת תדירות הפעמים שהם עלו בהרכב הפותח. הערה: אמנם לצורך חישוב מדד זה הסתכלנו גם על משחקים שייתכן שקרו אחרי המשחק הנוכחי, אך הסגל של קבוצה נבנה כמובן לפני המשחק, והפעולה שביצענו היא רק ניסיון לאמוד את טיב ההרכב הממוצע בו משתמשת הקבוצה. לכן לדעתנו, מבט על כל משחקי הקבוצה באותה עונה הוא מדויק יותר עבור חישוב רמת הקבוצה על הנייר מאשר מבט רק על משחקי עבר באותה עונה. נציין שרמות הקבוצות נעו בערכים שבין 65 ל-85.

- הרמה הממוצעת של עמדת השוער בקבוצה, וכנ"ל עבור הקבוצה היריבה. מדד זה חושב באופן זהה לחישוב המדד הקודם, רק שכעת הממוצע נעשה רק על פני השוערים ששיחקו בהרכב הקבוצה במשחקים השונים באותה עונה.

נציין כעת גם unmeasured confounders, כלומר confounders שלא נמצאים/לא ניתן לחשב במסד הנתונים שלרשותנו:

- מזג האוויר – במסד הנתונים לא מופיעים נתונים לגבי מזג האוויר בזמן המשחק. הוספה של נתונים אלו יצריכו איתור של המקום בו שוחק כל משחק (על סמך האצטדיון בו משחקת הקבוצה הביתית), איתור של השעות בהן התקיים המשחק, ואז הצלבה עם מסד נתונים של מזג אוויר המכיל נתונים עבור אותו מקום באותו הזמן. לא הצלחנו למצוא מסד נתונים אשר באמצעותו נוכל לבצע את תהליך האינטגרציה הנ"ל באמצעות קוד, ולכן בחרנו לוותר על נתון זה במגבלות הזמן שלרשותנו. עם זאת, נתונים שכן קיימים כמו השנה, החודש בשנה ומספר המחזור נותנים מידע על התקופה בשנה ולכן עשויים לפצות על חלק מהחוסר בנתוני מזג אוויר.
- זהות השופט – ייתכן וכל שופט משפיע אחרת על שטף המשחק ועל האופן בו הוא יתנהל, למשל רמת האגרסיביות בתיקולים שהוא יאפשר עלולה להשפיע על האופן בו הקבוצות יצליחו להחזיק בכדור ועל היכולת של הקבוצות להגיע למצבי בעיטה מסוכנים. איתור של השופט ששפט בכל משחק יהיה תהליך ידני גם כן על סמך הכלים שברשותנו, ולכן לא ביצענו אותו.
- כמות האוהדים באצטדיון – ייתכן כי לכמות הקהל באצטדיון ישנה השפעה על האופן בו יתפתח המשחק, למשל על ידי מתן "רוח גבית" לקבוצה הביתית או הרתעת קבוצת החוץ. עם זאת, נתונים על כמות האוהדים שהיו בכל משחק לא נמצאים במסד הנתונים שלנו, ולא מצאנו כיצד לבצע אינטגרציה ממקור אחר בצורה אוטומטית.
- חשיבות המשחק – אמנם יש לנו נתונים על רמת הקבוצות ועל מספר מחזור הליגה בו המשחק המתקיים, אך ייתכן כי ישנם מצבים בו המשחק מתקיים "לפרוטוקול בלבד", למשל משחק בין שתי קבוצות שיודעות כבר שאין להן סיכוי לשנות את מקומן בטבלה. במשחקים כאלו ייתכן ואופי המשחק יהיה אחרת מאשר במצב בו לקבוצות יש טעם להשיג ניצחון במשחק. אמנם מהנתונים שבידינו לא ניתן להסיק את חשיבות המשחק עבור כל קבוצה, אך כדי לנסות ולהימנע מהכללת משחקים חסרי חשיבות הסרנו את המשחקים משני המחזורים האחרונים בכל ליגה.

מדוע אפשר להגיע לשערוך קוזאלי?

בהרצאות דיברנו על כך שכדי לשערך את האפקט הקוזאלי על ידי covariate adjustment, יש להניח שמתקיימים ארבעה תנאים. נסביר כעת מדוע לדעתנו ניתן להניח שהם מתקיימים במסד הנתונים שלנו.

1. **SUTVA-** ה-potential outcomes שיש לכל יחידה (לשתי קבוצות במשחק נתון) לא משתנים מהטיפולים שאנו נותנים ליחידות אחרות. כל משחק מתקיים בנפרד, ומכל משחק אנו בוחרים להסתכל על קבוצה אחת בלבד. רמת השליטה בכדור של קבוצה במשחק אחר לא אמורה להשפיע על מה שיקרה לקבוצה במשחק הנוכחי בין אם תשלוט בכדור במחצית הראשונה ובין אם היא תבחר לשחק באופן פאסיבי.
בנוסף, אין וריאציות שונות של החלוקה שלנו לשתי קבוצות המחקר, אלא רק סוג חלוקה אחד, והוא לפי סף של אחוזי השליטה בכדור במחצית הראשונה.
על סמך נימוקים אלו נקבע שלדעתנו תנאי ה-SUTVA מתקיים.
2. **Consistency** – לכל קבוצה, אנחנו רואים את כמות הבעיטות למסגרת שבעטה הקבוצה במחצית השנייה בהתאם לדרך בה שלטה בכדור במחצית הראשונה, כלומר אנו רואים את ה-potential outcome שמתאים לקבוצת המחקר אליה משויכת הקבוצה. לכן הנחה זו מתקיימת.
3. **Ignorability** – הדאטה שנמצא במסד הנתונים שלנו הוא די עשיר וניתן להפיק ממנו confounders רבים, כמו למשל אופן הדרך בו קבוצות נוהגות לשחק, רמת סגל הקבוצות, יחסי הימורים שמרמזים גם הם על רמת שתי הקבוצות בכל משחק ועוד. עם זאת וכפי שציינו בסעיף הקודם, ישנם גם unmeasured confounders שלא ניתן להסיק מתוך מסד הנתונים שלנו. למרות זאת, אנו חושבים שה-confounders שנמצאים במסד הנתונים מספיקים לצורך ביצוע שיערוך קוזאלי, וננסה להסביר מה אנו עושים בכדי לצמצם השפעה אפשרית של ה-unmeasured confounders שציינו, או מדוע הם פחות משמעותיים לצורך השיערוך הקוזאלי:

- <u>מזג האוויר</u> – אמנם מזג האוויר עשוי להשפיע על האופן בו הקבוצות ישחקו ועל רמת המחץ ההתקפי שלהם, למשל אם הטמפ' נמוכה מאוד או שיורד גשם זלעפות, אך ההשפעה תהיה ברמה שווה על שתי הקבוצות ולכן ייתכן כי זה מבטל במעט את השפעת מזג האוויר. בנוסף, במסד הנתונים שלנו יש לכל משחק מידע עבור התאריך בו הוא התקיים, ושימוש בנתון כמו החודש בשנה בו התקיים המשחק יכול לתת בעקיפין מידע גם על טווח הטמפ' שסביר להניח שהיו בזמן המשחק.
- <u>זהות השופט</u> – גם במקרה זה, ההשפעה של שופט על המשחק היא לרוב מאוזנת בין שתי הקבוצות, ולכן ייתכן שדבר זה מבטל את רמת ההשפעה שלו ביחס לרמת השליטה בכדור של אחת הקבוצות לעומת האחרת למשל. בנוסף, בחרנו להסיר משחקים בהן נשלף כרטיס אדום לאחד השחקנים, ובכך אנו מסירים מקרים בהן ייתכן וההשפעה של שופט לרעת אחת הקבוצות הייתה ניכרת. בנוסף, הפיצ'רים שלנו מכילים מידע על כמות העבירות והכרטיסים הצהובים שהיו לכל קבוצה במחצית הראשונה. מידע זה מסייע להחליף את האינפורמציה על זהות השופט, שכן הוא יכול לרמז למשל על רמת האגרסיביות שהשופט מוכן לאפשר במשחק, והתדירות בה הוא שורק לעבירות.
- <u>כמות האוהדים באצטדיון</u> – אמנם אין לנו נתון מדויק על כמות האוהדים באצטדיון, אך אנו יודעים עבור כל משחק מי הייתה הקבוצה הביתית ואנו משתמשים במידע זה כ-confounder. סביר להניח כי תהיה השפעה במידה מסוימת לביתיות על הקבוצות המשחקות ללא קשר לכמות האוהדים באצטדיון, ולכן לדעתנו המידע שיש לנו מפצה על החסר.
- <u>חשיבות המשחק</u> – נשתמש בנתון של מספר המחזור בו התקיים המשחק כדי לסייע לצמצם את החוסר בכך שאיננו יודעים מהי רמת החשיבות כל משחק. נסיר עבור כל ליגה וכל עונה את המשחקים שהתקיימו בשני המחזורים האחרונים, בכדי לצמצם את כמות המשחקים בסט הנתונים שלנו שהתקיימו "לפרוטוקול", כלומר מבלי שניצחון במשחק היה עשוי לתרום לאחת הקבוצות. משחקים אלו יקרו לרוב בסוף העונה, כאשר מיקום הקבוצות בטבלה עשוי לא להשתנות או להשתנות בצורה

מינורית ללא תלות בתוצאת המשחק.
בנוסף, אנו נסיר משחקים בהם הפרש הכוחות בין הקבוצות גדול מדי. בכך נצמצם את האפשרות של משחקים לא תחרותיים שייתכן והקבוצות יתנהגו בהן באופן אחר מבדרך כלל. למשל, הקבוצה החלשה עשויה לוותר על משחקים כאלו מראש, או שהקבוצה החזקה תבחר לעלות בהרכב חסר מאוד ותשחק בצורה אחרת מאשר במשחק תחרותי יותר.

על סמך הנימוקים האלו, ועל סמך ה-confounders שכן נמדדים במסד הנתונים שלנו, אנו מניחים שתנאי ה-Ignorability מתקיים עבור מסד הנתונים.

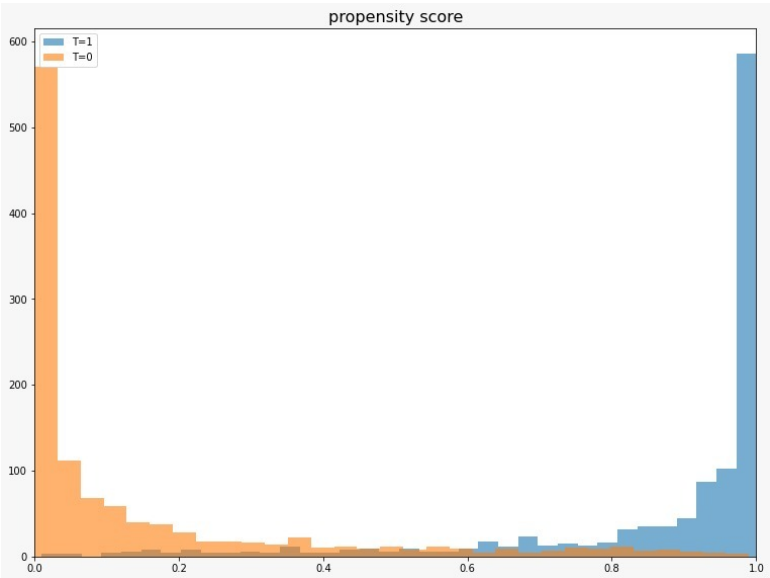
4. **Common Support (Overlap)** – בתנאי זה אנו דורשים כי:

$$p(T = t|X = x) > 0 \quad \forall t, x$$

כלומר שלכל תצפית תהיה הסתברות חיובית להיות בכל אחת משתי קבוצות המחקר.

כדי לבדוק האם סביר להניח שתנאי זה מתקיים, חישבנו לכל תצפית את ה- propensity score שלה, כלומר את $e(x) = p(T = 1|x)$. לאחר מכן, ביצענו Common support trimming (כפי שהוצג בתרגול 3), בכדי להגדיל את ה-overlap בין התצפיות שלנו.
בשיטה זו, אנו מסירים בכל קבוצת מחקר את כל התצפיות אשר אין בקבוצת המחקר השנייה תצפית עם propensity score קיצוני (לכיוון 0 או 1, תלוי בקבוצת המחקר) לפחות באותה המידה.

באיור 1 מוצגת היסטוגרמה של ה-propensity score עבור התצפיות השונות, מחולקות לפי קבוצות המחקר:



איור 1 – היסטוגרמת ערכי ה-propensity scores לפי קבוצות מחקר.

על סמך ההיסטוגרמה נראה כי ישנו overlap בין קבוצות המחקר, שכן בכל קבוצת מחקר ישנן תצפיות שערך ה-propensity שלהן יחסית קיצוני, למשל קרוב ל-1 אם אנו מסתכלים על הקבוצה $T = 0$.

בנוסף, בכך שאנו מסירים משחקים שבהם הפרשי הרמות בין הקבוצות גבוהים, ומשחקים שעשויים להיות משחקים לפרוטוקול, אנו מסירים מקרים שעלולים לפגוע ב-*overlap*, בין אם בגלל הפרשי הרמות, ובין אם לאחת הקבוצות לפחות אין טעם להשיג ניצחון במשחק.

על סמך הדברים שלעיל, אנו מניחים שתנאי ה-*Common Support* מתקיים עבור הנתונים בהם בחרנו להשתמש.

השיטות בהן השתמשנו

אנו ננסה לשערך את ה-ATE באמצעות שימוש בארבע שיטות שונות:

1. **Inverse Probability Weighting (IPW)** – בשיטה זו, ננסה למשקל את ה-outcome של כל תצפית באופן שיתקן את ההטיה של יחס הופעת תצפית בקבוצת המחקר בו היא הופיעה, זאת ביחס למקרה של RCT בה ההסתברות של כל תצפית להיות בכל אחת משתי קבוצות המחקר היא חצי. ההטיה הקיימת מתרחשת מכיוון שאנו משתמשים בנתונים מן העולם האמיתי, ולכן אין סיבה להניח שההשמה של קבוצות הכדורגל לשתי קבוצות המחקר היא מקרית.
כדי לנסות ולתקן הטיה זו, נאמן מודל ML שינסה לחזות לכל תצפית x את ההסתברות שהיא תהיה שייכת לקבוצת המחקר $T = 1$. באופן פורמלי, נחזה לכל תצפית x את $e(x) = p(T = 1|x)$, הנקרא propensity score. על ידי שימוש בערך ההופכי ל-propensity score, נוכל למשקל מחדש את ה-outcomes של התצפיות ובכך לנסות ולתקן את הטיית הופעתן ביחס לניסוי RCT.
שיערוך ה-ATE יתבצע על ידי הנוסחה הבאה:

$$\widehat{ATE} = \frac{1}{n}\sum_{i=1}^{n}\frac{y^i t^i}{\hat{e}^i} - \frac{1}{n}\sum_{i=1}^{n}\frac{y^i(1-t^i)}{1-\hat{e}^i}$$

כאשר:

- y^i – ה-outcome של התצפית ה-i.
- t^i – קבוצת המחקר אליה שייכת התצפית ה-i.
- $\hat{e}_i$ – ה-propensity score שנחזה עבור התצפית ה-i.
- n – מספר התצפיות.

אנו בחרנו להשתמש במודל Logistic Regression לצורך חיזוי ה-propensity score.

2. **S-Learner** – בשיטה זו אנו נאמן מודל f שיחזה את ה-potential outcome עבור כל קומבינציה של מאפייני התצפית וקבוצת המחקר, כלומר $\widehat{y^t} = f(x,t)$. בכדי לאמן את המודל, נשתמש בתצפיות שלרשותנו כאשר עבור כל תצפית i נשתמש בפיצ'רים שלה x^i ובקבוצת המחקר אליה היא שייכת t^i בתור הקלט, וב-outcome של התצפית y^i בתור המשתנה התלוי.
בשלב החיזוי, נשתמש במודל המאומן כדי לחזות עבור כל תצפית i את:
$$\widehat{y^1} = f(x^i, 1), \qquad \widehat{y^0} = f(x^i, 0)$$

כלומר ננסה לחזות עבור כל תצפית את שני ה-potential outcomes שלה. שיערוך ה-ATE יהיה ממוצע של הפרש חיזויים אלו על פני כל התצפיות שלרשותנו:

$$\widehat{ATE} = \frac{1}{n}\sum_{i=1}^{n}(f(x^i,1) - f(x^i,0))$$

אנו בחרנו להשתמש במודל Decision Tree Regressor בשביל לבנות את המודל f.

3. **T-Learner** – בשיטה זו נאמן שני מודלי חיזוי שונים, f_0, f_1, כאשר המודל f_0 יחזה את ה-potential outcome של כל תצפית לו הייתה שייכת לקבוצת המחקר $T = 0$ (ניתן לרשום כ- $\widehat{y^0} = f_0(x)$), והמודל f_1 יחזה את ה-potential outcome של כל תצפית לו הייתה שייכת לקבוצת המחקר $T = 1$ (ניתן לרשום כ- $\widehat{y^1} = f_1(x)$). נאמן כל אחד משני המודלים על ידי התצפיות ששייכות לקבוצת המחקר המתאימה לאותו מודל, כאשר הקלט עבור תצפית i יהיה הפיצ'רים של אותה תצפית x^i, והמשתנה התלוי יהיה ה-outcome של התצפית y^i.
בשלב החיזוי, נשתמש בשני המודלים המאומנים כדי לחזות עבור כל תצפית i את:
$$\widehat{y^1} = f_1(x^i), \qquad \widehat{y^0} = f_0(x^i)$$
כלומר ננסה לחזות עבור כל תצפית את שני ה-potential outcomes שלה. שיערוך ה-ATE יהיה הממוצע של הפרש חיזויים אלו על פני כל התצפיות שלרשותנו:
$$\widehat{ATE} = \frac{1}{n}\sum_{i=1}^{n}(f_1(x^i) - f_0(x^i))$$
אנו בחרנו להשתמש בשיטה זו במודל Decision Tree Regressor בשביל לבנות את המודלים f_0, f_1.

4. **1-NN Matching** – בשיטה זו אנו מנסים לדמות לכל תצפית מה היה קורה לו היא הייתה שייכת לקבוצת המחקר השנייה, על ידי חיפוש של התצפית הכי דומה לה מתוך קבוצת המחקר השנייה. על סמך מטריקת מרחק שנגדיר, נמצא עבור כל תצפית את התצפית שהכי קרובה לה מתוך התצפיות ששייכות לקבוצת המחקר השנייה, כאשר חישוב המרחקים והקרבה יכול להיעשות למשל על סמך סט הפיצ'רים של כל תצפית, או על סמך ה-propensity score. לאחר מכן, נחסר בין ה-outcomes של שתי התצפיות הללו, כאשר המחסר יהיה התצפית ששייכת לקבוצת המחקר $T = 1$, והמחוסר יהיה התצפית ששייכת לקבוצת המחקר $T = 0$. נבצע תהליך זה עבור כל תצפית שלרשותנו, ועל ידי מיצוע של כל התוצאות נקבל את השיערוך ל-ATE.
להלן פסאודו קוד של השיטה (מבוסס על שקופית 59 במצגת הרצאה 4 של הקורס):

Let $d(\cdot,\cdot)$ be a metric.
For each i = 1,…,n:
define $j(i) = \underset{j:\, s.t.\ t^j \neq t^i}{\operatorname{argmin}}\ d(x^j, x^i)$
if $t^i == 1$:
$\widehat{ITE}(i) = y^i - y^{j(i)}$
else:
$\widehat{ITE}(i) = y^{j(i)} - y^i$
$\widehat{ATE} = \frac{1}{n}\sum_{i=1}^{n}\widehat{ITE}(i)$

אנו בחרנו להשתמש בשיטה זו עם מטריקת המרחק האוקלידי. ננסה לבצע Matching ביחס לכל ה-covariates של התצפיות (נסמן כ-"matching_all"), ואחד נוסף על סמך ה-propensity scores (נסמן כ-"matching_propensity").

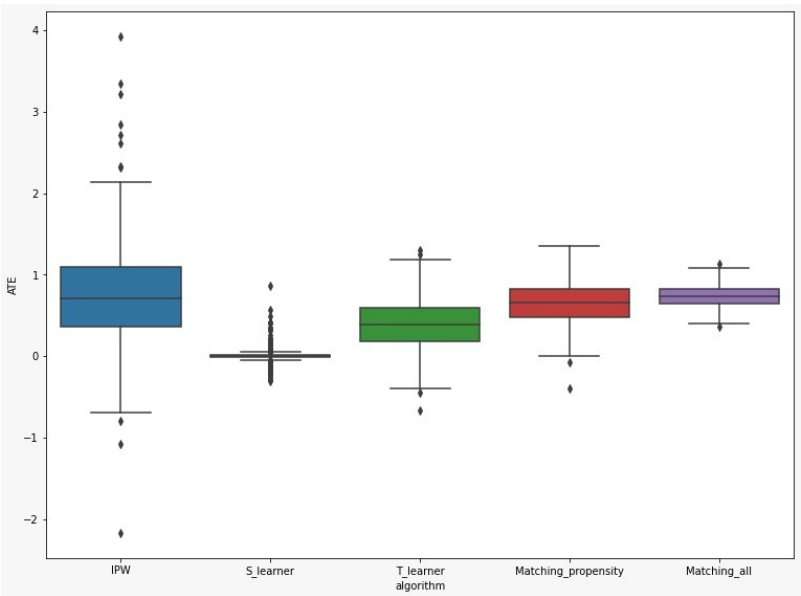
תוצאות וניתוחים נוספים

כאמור לעיל, הנתונים בהם השתמשנו הם של משחקים מתוך שש ליגות מרכזיות באירופה, כאשר אנו משתמשים רק במשחקים עם נתונים מלאים בהם:

- לא היו כרטיסים אדומים.
- אחת הקבוצות בהן החזיקה בכדור לפחות 60 אחוז מהזמן במחצית הראשונה.
- המשחקים לא נערכו באחד משני המחזורים האחרונים של הליגה באותה העונה.
- הפרשי הרמות בין הקבוצות לא היו גבוהים. חישוב הפרש הרמות נקבע על ידי ביצוע ערך מוחלט על הפרש רמת שחקני השדה שחישבנו עבור שתי הקבוצות. הסרנו משחקים בהם הפרש זה היה גדול או שווה ל-11 (כ-5 אחוזים מהמשחקים).

לאחר סינון המשחקים על פי הקריטריונים הנ"ל, חישוב ה-ATE התבצע על סמך 1168 משחקים. ביצענו את חישוב ה-ATE 500 פעמים, כאשר בכל פעם הגרלנו קומבינציה אחרת של קבוצות, קבוצה אחת מכל משחק. עשינו זאת על מנת להימנע מתלות בין התצפיות. מכיוון שבכל קומבינציה אחרת של הקבוצות תתקבל תוצאה שונה, הרצנו עבור מספר רב של קומבינציות ובחרנו לצייר boxplot עבור ערכי ה-ATE שהתקבלו.

איור 2 מציג את התוצאות ב-boxplot עבור כל אחד מהאלגוריתמים בהם השתמשנו לצורך חישוב ה-ATE:



איור 2 – boxplots המציגים את תוצאות הריצות עבור כל אלגוריתם.

באיור 2 ניתן לראות שלמעט אלגוריתם S-learner, עבור כל האלגוריתמים התקבל בממוצע ערך ATE חיובי, הנע בין 0.5 ל-1. אנו חושבים שהסיבה שעבור אלגוריתם S-learner התקבל בממוצע ATE הקרוב ל-0 היא שהמודל אותו אימנו בשיטה זו למד משקל נמוך מאוד לפיצ'ר

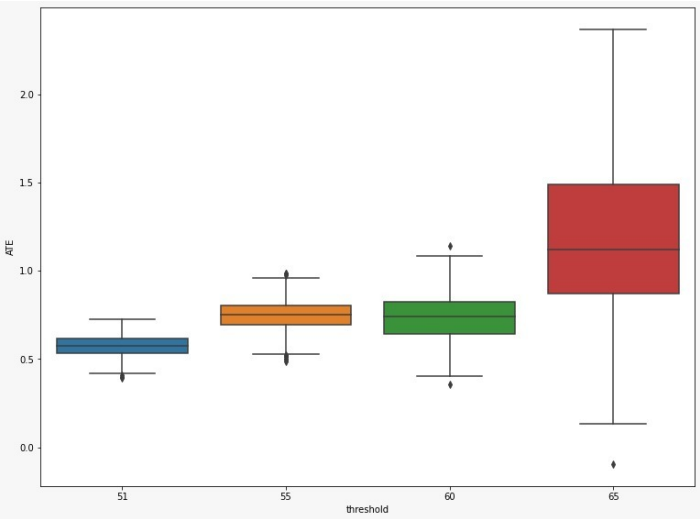
של קבוצת המחקר (T), אולי עקב מספר רב של משתנים מסבירים. כתוצאה מכך, הפלט שהוא החזיר עבור כל תצפית היה כמעט זהה, בין אם נתנו לתצפית זו T=1 או T=0.

בנוסף, ניתן לראות כי עבור אלגוריתם ה-IPW התקבלה שונות יחסית גבוהה לשאר האלגוריתמים. אנו משערים שהסיבה לכך היא ערכי ה-propensity score שנחזו לכל תצפית. באיור 1 ראינו שרוב ערכי ה-propensity score היו יחסית גבוהים עבור התצפיות ששייכות ל-T=1, ונמוכים יחסית עבור התצפיות ששייכות ל-T=0. ייתכן שזה משפיע על יציבות התוצאה שאלגוריתם ה-IPW יחזיר עבור בחירת קומבינציה שונה של קבוצות מכל המשחקים.

נתאר כעת ניתוחים נוספים שערכנו, ואת התוצאות שקיבלנו עבורן:

1. בדקנו כיצד ישתנה ערך ה-ATE עבור חלוקות שונות לקבוצות המחקר. החלוקה הבינארית בה השתמשנו במהלך המחקר הייתה האם הקבוצה הייתה זו ששלטה בכדור לפחות 60 אחוז מזמן המחצית הראשונה, או ששלטה בכדור לכל היותר 40 אחוז מהזמן במחצית הראשונה. כעת, ננסה לשנות את ההגדרה של "שליטה בכדור" לאחוזים שונים של החזקה בכדור מסך זמן המחצית הראשונה, ונראה את ההשפעה שלהן על תוצאות ה-ATE. הערכים השונים של אחוזים שלפיהם ניסינו לחלק לקבוצות מחקר הם 51, 55, 60, 65.

איור 3 מציג עבור כל אחד מהערכים הנ"ל boxplot של תוצאות ערכי ה-ATE על סמך 500 סימולציות של קומבינציות של קבוצות. האלגוריתם בו השתמשנו לחישוב ה-ATE הוא matching_all, כלומר 1-NN matching על סמך כל ה-covariates.

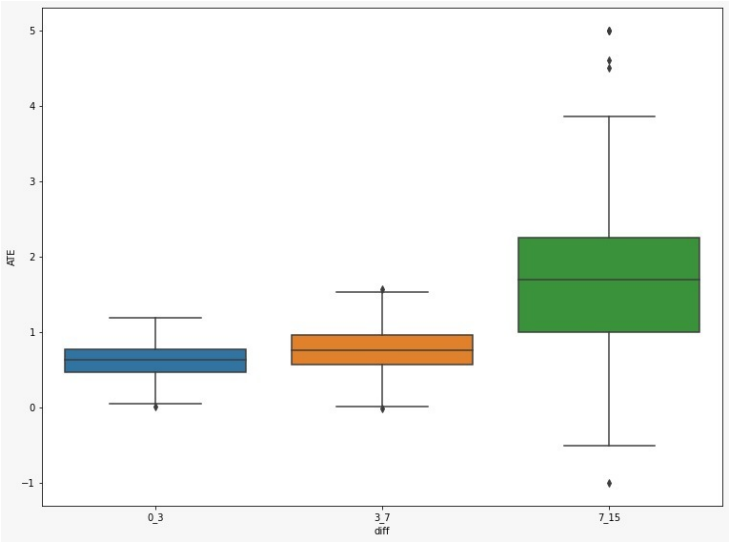


איור 3 - boxplots המציגים את תוצאות הריצות עבור כל סף חלוקה לקבוצות מחקר.

באיור 3 ניתן לראות שעבור כל ערכי הסף לחלוקה לקבוצות מחקר קיבלנו בממוצע ערך חיובי של ATE, הנע בין כ-0.6 עד לכ-1.2. ניתן גם לראות שהמגמה היא שככל שדרשנו שהקבוצה ה"שולטת" תהיה בעלת אחוזים גבוהים יותר של החזקה בכדור במחצית הראשונה, כך הערך הממוצע של ה-ATE גדל. בנוסף, ניתן לראות שעבור סף החלוקה של 65 אחוזים השונות בין ריצות הייתה די גבוהה (מתבטא בטווח רחב של ה-boxplot). לדעתנו, הסיבה לכך היא מיעוט יחסי בכמות המשחקים שענו על הסף הזה (כמה מאות בלבד), ולכן השונות בין קומבינציות של בחירת קבוצה אחת מכל משחק גבוהה יותר.

2. לאחר התוצאות הללו, רצינו לנסות ולבדוק את רמת ההשפעה של הפרשי הרמות בין הקבוצות על מדד ה-ATE שהגדרנו. בתוצאות הראשיות הסרנו משחקים עם הפרש רמות גבוה מ-11, כאשר הפרש הרמות חושב על סמך רמת שחקני השדה שחישבנו לכל קבוצה. כעת, רצינו לבדוק מה יקרה לו נשנה את המדד, ונסתכל על משחקים עם הפרשי רמות נמוכים/ גבוהים יותר.
כדי לבצע זאת, חילקנו את המשחקים שלרשותנו, לאחר שביצענו את הסינונים שלא נוגעים להפרשי הרמות, לשלוש קבוצות הדומות בגודלן: משחקים בהם הפרש הרמות בין הקבוצות היה בטווח של 0-3, משחקים בהם הפרש הרמות היה בטווח של 3-7, ומשחקים בהם הפרש הרמות היה 7 ומעלה.

עבור כל אחת משלוש הקבוצות הללו, השתמשנו באלגוריתם matching_all בכדי לייצר boxplot על סמך 500 קומבינציות של בחירת קבוצה מכל משחק. התוצאות מוצגות באיור 4:

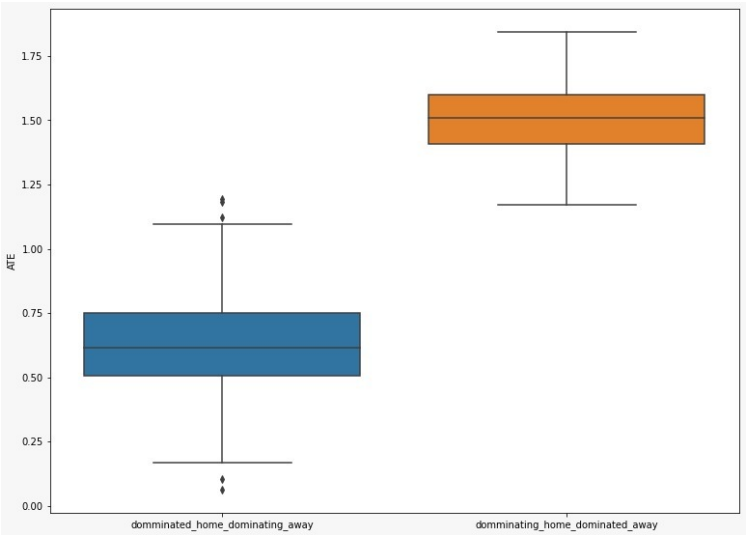


איור 4 - boxplots המציגים את תוצאות הריצות עבור כל טווח של הפרשי רמות.

ניתן לראות שעבור כל טווחי הפרשי הרמות, ממוצע ה-ATE שהתקבל היה חיובי (הנמוך ביותר שהתקבל – כ-0.6; הגבוה ביותר שהתקבל – כ-1.6). בנוסף, ניתן לראות עלייה בממוצע ה-ATE ככל שטווח הפרשי הרמות הכיל ערכים גבוהים יותר. ניתן גם לראות שהשונות בין ריצות הייתה גבוהה יותר ככל שטווח הפרשי הרמות הכיל ערכים גבוהים יותר. תוצאה זו קצת הפתיעה אותנו, שכן כמות המשחקים בכל אחת משלוש הקבוצות היא די דומה.

3. ניתוח נוסף שביצענו הוא של השוואה בין ביצועי הקבוצות ששולטות בכדור במחצית הראשונה כאשר הן הקבוצה הביתית, לעומת המקרה בו הן קבוצת החוץ. חישבנו את ה-ATE על סמך המשחקים בהם הקבוצה הביתית הייתה הקבוצה שהחזיקה בכדור לפחות 60 אחוז מהזמן במחצית הראשונה, וביצענו את אותו התהליך פעם נוספת עבור המשחקים בהם אלו היו קבוצות החוץ ששלטו בכדור. רצינו לראות האם יש הבדל בין יכולת הקבוצות ה"שולטות" כאשר הן הקבוצות הביתיות לעומת המקרה בו הן קבוצת החוץ.

איור 5 מציג עבור כל אחד משתי האפשרויות שלעיל boxplot של תוצאות ערכי ה-ATE על סמך 500 סימולציות של קומבינציות של קבוצות. האלגוריתם בו השתמשנו לחישוב ה-ATE הוא matching_all.



איור 5 - boxplots המציגים את תוצאות הריצות עבור משחקים בהם הקבוצה השולטת הייתה קבוצת הבית/החוץ.

מן התוצאות שבאיור 5 ניתן לראות פער יחסית משמעותי בין ממוצע ה-ATE עבור משחקים בהם הקבוצה השולטת הייתה קבוצת הבית (כ-1.5), לעומת ממוצע ה-ATE כאשר הקבוצה השולטת שיחקה בחוץ (כ-0.62). מן האיור גם ניתן לראות כי השונות בין ריצות עבור המקרה בו הקבוצות השולטות היו הקבוצה הביתית דומה לשונות בין ריצות עבור המקרה בו הקבוצות השולטות היו קבוצת החוץ.

חולשות אפשריות

נציין כעת כמה חולשות אפשריות שעלולות להשפיע על אמינות התוצאות שקיבלנו:

- כמו שציינו, ישנם unmeasured confounders שונים (למשל מזג אוויר) שעלולים להשפיע על אמינות התוצאות שלנו ועל רמת הדיוק שלהן. עם זאת, ניסינו לנקוט באמצעים שונים על מנת למנוע כמה שניתן את ההיתכנות של מצבים כאלו, כפי שפירטנו לעיל.
- הנתונים בהם השתמשנו על הרמה של שחקני הקבוצות, ועל השיטות בהן הקבוצות משחקות אינם מעודכנים ליום המשחק עצמו. לרוב אלו נתונים שעודכנו בתחילת/באמצע העונה, ולכן ייתכן פער של מספר חודשים בין משחק של הקבוצה לבין הנתונים שיש לנו אודות שיטת המשחק של הקבוצה והרמה של שחקניה. עם זאת, נציין שמאפיינים כמו רמת השחקנים לא נוטים להשתנות באופן דרסטי בטווחי זמן של חודשים, ולכן נתונים אלו עדיין די עדכניים.
- הנתונים על הרמה של שחקני הקבוצות, ועל השיטות בהן הקבוצות השונות משחקות הם כמובן נתונים סובייקטיביים שהוגדרו על ידי בני אדם, ולא מדדים אבסולוטיים. הנתונים נאספו מתוך משחק המחשב FIFA, וייתכן שהם אינם משקפים תמיד את רמת השחקן/הקבוצה במציאות. על פי אחד מהמפיקים של החברה בה נוצר המשחק[4], נתוני השחקן מושפעים גם מרמת הליגה בה הוא משחק ומהשיטה בה הקבוצה שלו משחקת, כך שלעיתים גם קריטריונים נוספים מלבד איכות השחקן נלקחים בחשבון. עם זאת, על פי אותו מפיק, תהליך דירוג השחקנים הוא תהליך מאסיבי הכולל מעל 9000 אנשי מקצוע ואוהדים המדרגים על סמך נתונים וקטעי וידאו של השחקנים. לכן, אף על פי

שייתכנו אי דיוקים ביחס למציאות, דירוג השחקנים והקבוצות על סמך משחק המחשב FIFA הוא דירוג מקצועי ומבוסס.

- למרות שבסט הנתונים המקורי היו כ-5,500 משחקים עם נתונים מלאים, לאחר תהליכי הסינון השונים שביצענו נותרנו עם כמות קטנה בהרבה של משחקים. למשל, עבור הניתוח המרכזי שלנו השתמשנו ב-1168 משחקים. בנוסף, לאחר תהליך הסינון נותרנו גם עם ייצוג לא מאוזן של הקבוצות והליגות השונות במסד הנתונים. על אף שהתוצאות שמצאנו התבססו על מספר משחקים די גדול למרות הסינון, אנו חושבים שעל סמך מסד נתונים גדול יותר המכיל כמות רבה יותר של משחקים נוכל להשיג תוצאות יותר מדויקות ואמינות.

מסקנות ודיון:

במהלך המחקר בדקנו את ההשפעה של יחס הקבוצה בנוגע לשליטה בכדור במחצית הראשונה על יכולת הקבוצה במחצית השנייה, כאשר את היכולת מדדנו באמצעות כמות הבעיטות למסגרת שהיו לקבוצה במחצית השנייה. על סמך שיטות שונות שלמדנו בהרצאה לאמידת ה-ATE, קיבלנו כי הקבוצה שהחזיקה בכדור בצורה דומיננטית במחצית הראשונה בועטת יותר למסגרת במחצית השנייה (בתוחלת) מאשר הקבוצה שהחזיקה פחות בכדור בזמן המחצית הראשונה (איור 2). למעט אלגוריתם S-learner, שלהערכתנו לא הצליח ללמוד את השפעת השייכות לאחד משתי קבוצות המחקר, כל שאר האלגוריתמים שניסינו שיערכו את ה- ATE להיות בין 0.5 ל-1. הסכמה יחסית זו בין האלגוריתמים מחזקת לדעתנו את אמיתות התוצאות, כך שנראה כי לשליטה בכדור במהלך במחצית הראשונה יש השפעה חיובית על היכולת של הקבוצה השולטת במחצית השנייה ביחס לקבוצה נגדה היא משחקת.

לאחר קבלת תוצאות אלו, ניסינו לראות כיצד שיערוך ה-ATE היה משתנה לו היינו מגדירים את מונח ה"שליטה בכדור" בצורה שונה, למשל 51 אחוזי החזקה בכדור במקום 60. קיבלנו כי ככל שדרשנו סף גבוה יותר של אחוזי החזקה בכדור, כך ממוצע ה-ATE היה גבוה יותר. עם זאת, שמנו לב (איור 3) שממוצע ה-ATE על סמך הריצות השונות שערכנו לכל ערך סף היה דומה מאוד עבור ערך סף של 55 אחוזי החזקה בכדור ושל 60 אחוזי בכדור. נראה כי לעומת הפער שקיבלנו בין 51 אחוזי החזקה ל-55 אחוזי החזקה, או הפער שקיבלנו בין 60 אחוזי החזקה ל-65 אחוזי החזקה, המעבר בין 55 ל-60 איננו משפיע באופן שמשנה את הפער בין יכולת הקבוצות במחצית השנייה.

בחרנו לנתח גם את המשחקים בסט הנתונים לפי רמת הפער האיכותי על הנייר בין הקבוצות המתחרות, ושיערכנו את ה-ATE עבור רמות שונות של טווחי פערי איכות בין הקבוצות. קיבלנו (איור 4) כי ערך ה-ATE גדל ככל שרמת האיכות בין הקבוצות הייתה גדולה יותר. נראה אם כן שכמו שהיינו מצפים, לפער ביכולת במחצית השנייה בין הקבוצה ששלטה בכדור במחצית הראשונה לבין זו שמיעטה להחזיק בכדור יש קשר גם לפערי האיכות בין הקבוצות מבחינת טיב הרכב השחקנים.

לבסוף, בחנו את ההבדל בערך ה-ATE שהתקבל בין משחקים בהם הקבוצה שהחזיקה בכדור במחצית הראשונה הייתה הקבוצה המארחת, לבין המקרים בהם היא הייתה קבוצת החוץ. קיבלנו (איור 5) פער די משמעותי בין ערכי ה-ATE עבור הבדיקה הזו, כאשר ממוצע ה-ATE שהתקבל על סמך ריצות שחושבו על סמך משחקים בהם הקבוצה שהחזיקה בכדור במחצית הראשונה הייתה קבוצת הבית היה יותר גדול יותר מפי שניים מממוצע ה-ATE שהתקבל על סמך ריצות שחושבו על סמך משחקים בהם הקבוצה שהחזיקה בכדור במחצית הראשונה הייתה קבוצת החוץ. נראה אם כן שלפער ביכולת במחצית השנייה בין הקבוצה

ששלטה בכדור במחצית הראשונה לבין זו שמיעטה להחזיק בכדור יש קשר לביתיות של הקבוצות. ייתכן כי לקבוצה שמחזיקה בכדור יותר נוח לבנות התקפות וליזום אל מול הקהל הביתי שלה שמעודד אותה כאשר היא מחזיקה בכדור, ולא אל מול קהל עוין שעלול להוציא אותה מריכוז ולפגע במחץ ההתקפי של ניסיונות הקבוצה. ייתכן גם כי הקבוצה שאיננה מחזיקה בכדור מצליחה להתגונן טוב יותר כאשר הקהל הביתי שלה תומך בה. עידוד הקהל עשוי לתת לשחקנים מרץ להמשיך להתאמץ ולנסות ולסכל את התקפות היריבה, גם כאשר הם עצמם לא מנסים או לא מצליחים להחזיק בכדור.

לסיכום, מהתוצאות שקיבלנו עולה כי ישנה השפעה חיובית לשליטה בכדור במחצית הראשונה על יכולת הקבוצה להגיע למצבים איכותיים במחצית השנייה. המלצתנו לקבוצות כדורגל אם כן היא לנסות ולקדם שיטת משחק של החזקה בכדור והכתבת קצב המשחק, כיוון שעל פי המחקר שלנו הדבר יסייע להגדלת פער היכולת של הקבוצה לעומת יריבתה בחלקיו המתקדמים של המשחק.

ביבליוגרפיה

1. European Soccer Database:
https://www.kaggle.com/hugomathien/soccer
2. Supplementary Database:
https://www.kaggle.com/jiezi2004/soccer
3. המאפיינים השונים שיש בסט הנתונים על כל קבוצה ומשמעותם:
https://www.fifplay.com/fifa-17-tactics/
4. כתבה על הדרך בה נקבע דירוג שחקני הכדורגל ב-FIFA:
https://www.vg247.com/2016/09/27/how-ea-calculates-fifa-17-player-ratings/