**מבוא**

כמו לרבים ממשחקי הספורט, גם סביב משחק הכדורגל התפתחה תרבות הימורים ענפה. מהימורים פשוטים על זהות המנצחת ועד להימורים יותר מסובכים כמו מי יבקיע ראשון או כמה אחוזי החזקת כדור יהיו לקבוצה מסוימת, ואפילו הימורים אזוטריים כמו מה סכום מספר החולצות של השחקנים שיכבשו גול או האם איזו קבוצה תבעט את כדור הקרן האחרון. המשך קיומם של סוכנויות ההימורים הם העדות החזקה ביותר להיותם הצד המרוויח מכל הסיפור. הצלחות קטנות של אנשים מתגמדות למול סכומי הכסף שמגלגלים אותן סוכנויות. הרבה אנשים סבורים שחוכמתם עומדת להם בכדי לנצח את משחק ההימורים ולהרוויח. האם הם צודקים? האם ידע בכדורגל משפר את סיכויי הזכייה?

במאמר של דנון ושות'[1] נטען כי אין לידע ובקיאות במשחק השפעה משמעותית על אחוזי הפגיעה של מהמרים. לצורך הניסוי נלקחו שלוש קבוצות של מהמרים כל אחת עם רמת הבנה שונה בכדורגל: מהמרים, חובבים וחסרי ניסיון. ביצועי שלוש הקבוצות הייתה זהה.

מצד שני מהמרי ספורט רבים משוכנעים שידע עוזר להם להרוויח, חלקם טוענים לרווח קבוע ומתמשך, ובנוסף קיימים אתרים שנותנים עצות למהמרים שפועלים בהצלחה מסוימת כבר שנים רבות. יש לציין שסוכנויות ההימורים לא נותנות יחס זכיה שווה עבור שתי הקבוצות במשחק, אלא מתעדפים את אחת הקבוצות ונותנים לה יחס נמוך יותר - הקבוצה שסוכנות ההימורים חושבת שהיא הפייבוריטית. כלומר - בקביעת היחסים על ידי סוכנות ההימורים קיים, מלבד האלמנט של הבנה בכדורגל, גם גורם הסתברותי וייתכן מאוד שבקביעתו משולבים בין היתר היבטי בינה מלאכותית.

נרצה לבחון טענה זו של דנון ושות' אל מול הטענות הגורסות אחרת. אנו רוצים להוכיח כי ידע מוקדם ויכולת ניתוח יכולים לסייע בזכייה (אם כי במקרה זה אנו עומדים להשתמש ביכולות החישוב של מחשב). כמו כן, נראה כי אחוזי הפגיעה בתוצאה הנכונה אינם הפרמטר היחיד בו יש להתחשב. קיימים מחקרים[2] שמטרתם לחזות בהסתברות גבוהה מי הקבוצה שתנצח - זה נושא מעניין אבל לא הנושא שבחרנו להתמקד בו. לא קשה להגיע לאחוזי דיוק טובים למדי לגבי איזו קבוצה תנצח - קל יחסית להמר תמיד על הקבוצה הפייבוריטית. העניין הוא שהימור בשיטה כזאת יוביל כנראה להפסד בטווח הארוך כפי שיוסבר בהמשך. החוכמה אינה להמר על הקבוצה שהכי סביר שתנצח אלא על הקבוצה שיחס הזכייה עבורה משתלם ביחס להערכה שלנו את סיכויי הניצחון שלה.

הערה - חשוב לציין שכוונתנו היא אקדמית בלבד ואין בה בכדי לעודד אנשים להמר, או להמליץ להמר בשיטה מסוימת על חשבון אחרות או לתת טיפים למהמר, ואנחנו לא לוקחים אחריות על כל שימוש במסקנות פרויקט זה למטרות שאינן אקדמיות נטו.

הערה נוספת: בפרסומות ל"ווינר" הישראלי מופיעה הערה שהזכייה תלויה במזל בלבד-רמיזה לכך שאין לצפות לאלמנט משמעותי של יכולת.

עוד יש לציין שלצורך הפרויקט לא בוצע שום הימור, אמנם השתמשו בנתונים של סוכנות הימורים בריטית שאינה זמינה למשחק בישראל(לא חוקי בישראל להמר באתרים מחו"ל) אבל לא נעשה שום הימור אמיתי - הכול בסימולציה שמבוצעת אצלנו במחשב.

בתמונה-"פול התמנון"-שזכה לפרסום במונדיאל 2010 על ידי אחוזי הצלחה גבוהים מאוד ב"חיזוי" איזו קבוצה תנצח, על ידי הליכה לקופסת האוכל המתאימה. למרות שכמובן שתוצאות אלה מקריות לחלוטין (בהנחה שלא שמו אוכל שהוא מעדיף באחת הקופסאות, ולטענת הבעלים שלו-שמו אוכל זהה בשתי הקופסאות) עדיין ההצלחה של התמנון הפכה אותו לאגדה בקרב חובבי כדורגל ומהמרי ספורט. בתמונה זו הוא "מנבא"(נכונה) שספרד תנצח את הולנד בגמר מונדיאל 2010. לצערנו התמנון כבר מת ולא יוכל להוות אצלנו תכונה עבור משחק כלשהו...



**תיאור הפתרון המוצע לבעיה**

הרעיון הוא מימוש אלגוריתם למידה שיעריך, על סמך נתוני עבר, את ההסתברות לכל אחת משלוש התוצאות האפשריות למשחק: ניצחון לקבוצת הבית, לקבוצת החוץ או תיקו. על בסיס ההסתברות שנקבל נממש אסטרטגיית הימור בתקווה שתמקסם את הרווח.

העדפנו להתמקד בהימור הבסיסי יחסית ולא דברים יותר אזוטריים יותר כמו שצוינו למעלה, משום שהימורים אלה נפוצים בכל סוכנויות ההימורים, כמעט לכל משחק שקיים, ולרוב התוחלת שלהם טובה יותר, בנוסף אלה נתונים שלהערכתנו יש יותר אלמנט של היסק לוגי מאשר מזל נטו מול נתון כמו מספרי החלוצות של הכובשים.

**הגדרות**

האובייקט אותו אנו רוצים לסווג הוא **משחק** בין קבוצת בית לקבוצת חוץ. תכונות המשחק הן תכונות הקבוצות המשתתפות בו בהצגה כזו או אחרת (יבואר בהמשך).

באופן פורמלי:

פונקציה כלשהי.

כאשר מתקיים:

**תכונות**

התכונות של משחק הן בעיקרן תכונות של הקבוצות המשתתפות במשחק.

תכונות גלובליות כמו מזג אוויר או שופט, הן בעיתיות משתי סיבות: האחת קשה להשיג נתונים על תכונות אלו אחורה בזמן. סיבה נוספת היא שהאובייקט הוא משחק. האלגוריתם לא יודע להבדיל בין ריאל מדריד לברצלונה ולא יודע להעריך איך שחקניה מתמודדים למשל במזג אוויר קשה. אנו מניחים שהשפעתם של גורמים גלובליים אלה על הביצועים של שתי הקבוצות קרובה להיות זהה.

כמו כן, בבחירת התכונות התייחסנו לרמת הקבוצה ולא ירדנו לרזולוציית השחקנים. בחירה זו נובעת משלוש סיבות עיקריות: ראשית, לא תמיד קל להשיג מידע על כל השחקנים בכל קבוצה. שחקנים רבים הם אלמונים ואין עליהם הרבה מידע. שנית, איכות השחקנים בדרך כלל מגולמת באיכות הקבוצה. הסיבה השלישית היא שגם אם שחקן משחק בקבוצה מסוימת זה לא אומר שהוא ישתתף במשחק. בכל קבוצה משחקים 11 שחקנים בלבד מתוך 18. אף שהרכב הפתיחה בדרך כלל קבוע, קשה להעריך אלו שחקנים ישתתפו במשחק בעקבות אירועים לא צפויים כמו הרחקות, פציעות וחילופים. גם קשה להעריך השפעה של שחקן בודד בפני עצמו אלא יש חשיבות לשיטת המשחק, שיתוף פעולה עם שחקנים אחרים, כושר השחקן(אם הוא חזר מפציעה) ועוד. עם זאת, קיימת עבודה מעניינת בתחום, אשר מסתמכת אך ורק על נתוני השחקנים שמגיעה לרווח של למשחק בממוצע[2].

את התכונות שנבחרו ניתן לחלק לשלושה סוגים:

1. **נתונים סטטיסטיים**:

עבור כל קבוצה אספנו נתונים סטטיסטיים במשך העונה הנוכחית וקומבינציות שלהן: שערים למשחק, נקודות בטבלה, הפרש שערים, כרטיסים, בעיטות למסגרת, מספרים משחקים ללא שערים\ללא ספיגת שערים נתונים דומים מחמשת המשחקים האחרונים ועוד...

הרשימה המלאה מצורפת בנספחים.

סך כל התכונות בקטגוריה זו עבור כל אחת מהקבוצות הוא 100.

1. דירוגים כלליים:

השתמשנו בשני דירוגים חיצוניים:

1. **Elo rating** או בעברית **מד כושר**–שיטת דירוג שהתפתחה במקור עבור שחמט והורחבה לתחום הכדורגל ומשחקים תחרותיים רבים נוספים-וידוע כמדד חזק עבור חיזוי מי מנצח. עיקרון השיטה הוא שניצחון על קבוצה טובה זה יותר "נחשב" מאשר ניצחון על קבוצה נחותה, והפסד לקבוצה נחותה נחשב יותר גרוע מאשר הפסד לקבוצה טובה.

באתר שממנו אנחנו לוקחים את דירוגי הelo, שמתוחזקים קבוע מתוצאות של שנים רבות(על ידי סקריפט משלהם) מכל הליגות הגדולות, נלקחים בחשבון הפרמטרים הבאים בנוסף לדירוג הקלאסי:

* הפרש השערים במשחק-ככל שניצחון הוא בהפרש גדול יותר ככה הוא משפיע יותר
* משחקי בית וחוץ-ניצחון בית נחשב פחות מאשר ניצחון חוץ, שכן יותר קל לנצח בבית.

פירוט מלא של השיטה מופיע באתר הבא:

<http://clubelo.com/System>

\*יש לציין שהאתר לוקח בחשבון גם משחקים במפעלים שאנחנו לא מסתכלים עליהם-כמו ליגת האלופות וגביעים.

דירוג זה הוא דינמי ומשתנה לאחר כל משחק. לכל משחק יש להתאים את הדירוג המתאים של כל קבוצה לפי תאריך.

למרות שיש שיטה להמיר תוצאות אלה לחיזוי הסתברויות החלטנו שלא לעשות זאת אלא להתחשב בדירוג הelo כתכונה נוספת עבור כל משחק.

הדירוג המקסימלי האפשרי לקבוצה הוא קצת יותר מ 2000 (ולרוב 2000 זה חסם עליון), הנתון הנמוך ביותר שיש לנו נתונים עליו הוא קצת מעל 1000 ולכן החלטנו שכל קבוצה שאין עליה משום מה נתונים עבור משחק נתון אז ניתן לה ערך של 1000 - בגלל שאנחנו מניחים שקבוצות שאין עליהן נתונים יהיו כנראה קבוצות שעלו מליגה נמוכה ואינן קבוצות חזקות כל כך.

1. דירוג קבוצות של משחקי וידאו (fifa) – זהו מדד חצי שנתי הניתן לכל קבוצה ע"י יצרני משחקי הוידאו. אנו מניחים שמטרת היצרנים היא לתת חווית משחק קרובה למציאות ולכן הדירוג שלהם מקצועי ומבוסס.

מחשב המחשב FIFA הוא משחק מחשב שמהווה סוג של סימולטור כדורגל, אהוב מאוד על חובבי כדורגל ומשחקי מחשב בעולם. במשחק מיוצגות הקבוצות האמיתיות, עם שחקנים שאמורים לשקף את השחקנים מהעולם האמיתי. הייצוג אינו רק במראה והשם של השחקנים והקבוצות אלא גם בנתונים שקובעים את יכולת השחקן במשחק-למשל מהירות, גובה, יכולת הבקעה, יכולת הגנה וכו'.

נתונים אלה נקבעים על ידי מומחי כדורגל מרחבי העולם, וזוכים לא אחת לתגובות מצד השחקנים האמיתיים שמשתמשים בזה כסוג של הערכת מומחים עבורם. לכן אנחנו משערים שניתן להיעזר בציונים אלה כדי להסיק איזו קבוצה טובה יותר.

החלטנו שלא להיכנס לרזולוציה של שחקנים בודדים אלא להסתכל על ממוצע הציון של הקבוצה(של ההרכב הפותח), מסיבות דומות לאלו שהוזכרו קודם לגבי למה אנחנו לא מסתכלים על ביצועי שחקנים בודדים.

1. **"ראש בראש"** : היסטוריית תוצאות המשחקים בין הקבוצות הספציפיות במשחק. קטגוריה זו כוללת שלוש תכונות: הפרש ניצחונות בין הקבוצות, הפרש שערים בין הקבוצות והפרש ניצחונות בין הקבוצות כאשר ניצחונות רחוקים יותר בזמן קיבלו משקל נמוך יותר.

ההיגיון אומר שמה שחוזה הכי טוב משחק בין שתי קבוצות הוא משחק שכבר היה בין הקבוצות-וככל שהתנאים יותר דומים אז הוא יחזה טוב יותר. יש לציין שבליגות הגדולות יש שני סיבובים בלבד בכל עונה-משחק בית ומשחק חוץ, ככה שלא ייתכן שבעונה עונה יהיה משחק נוסף שמתקיים עם אותם תנאים בדיוק(במסגרת ליגה-התעלמנו מגביעים וטורנירים כלל אירופיים) לצערנו...

הערה: ניתן היה להתחשב ביחסים שהתקבלו מחברת ההימורים בתור תכונה של המשחק, או ביחסים שמספקות חברות אחרות-שכן זה מידע שגלוי בפני מהמר, ומהמרים רבים מתחשבים בו. עם זאת די ברור שזה פרמטר מאוד מוצלח מבחינת חיזוי ההסתברויות שכן חברות ההימורים צריכות לחזות טוב את הסיכויים... לכן כנראה שתכונות אלה היו משתלטות על כל התכונות האחרות וקובעות כמעט לבדן את התוצאה.

לפיכך ראינו בזה נתון לא מעניין שיפגע ברמת העניין של הפרויקט ולא ישפר אותה.

**איסוף מידע**

את המידע הסטטיסטי חילצנו מהמידע הקיים באתר football-date[2] . אתר זה מכיל רשימות של משחקים עם הנתונים הבאים: מספר שערים, בעיטות לכיוון השער, בעיטות למסגרת, כרטיסים, ויחסי ההימורים של מספר סוכנויות הימורים מובילות בעולם\*. עבור כל משחק, חישבנו לכל אחת מהקבוצות את ביצועיה לאורך העונה הנוכחית עד לאותו משחק, תוך שאנו מתחשבים בעובדה שאחת משחקת כעת בבית ואחת בחוץ.

\*יש לציין שיחסי ההימורים משתנים מספר פעמים לפני המשחק –אנחנו מניחים שמדובר ביחס העדכני ביותר שפורסם זמן קצר מאוד לפני תחילת המשחק-אבל זה לא בהכרח המצב. בכל מקרה-זה כן יחס שבשלב מסוים היה זמין למהמרים, וסביר להניח שהנתונים נאספו בצורה אחידה עבור כל המשחקים והליגות.

את נתוני ה-**elo**  השגנו מהאתר[3]: <http://clubelo.com>. מיינו אותם לפי תאריכים וקבוצות ושייכנו אותם למשחקים הרלוונטיים.

ציוני משחק ה-**fifa** נלקחו מהאתר .

המידע שנאסף הוא עבור חמש ליגות כדורגל מהבכירות בעולם: הליגה האנגלית, הצרפתית, הגרמנית, הספרדית והאיטלקית, הליגות הראשונה והשנייה. בליגות אלה מהימנות תוצאות המשחקים גבוהה ביותר ותופעות פסולות של שוחד פחות נפוצות בהן, וגם איסוף המידע עליהן יותר אמין ומסודר מאשר ליגות אחרות, וציוני שחקנים של קבוצות כאלה קיימים במשחק FIFA. נאסף מידע על 14 עונות: מעונה 2006/2007 ועד לעונה האחרונה 2019/2020. בקבוצת האימון נבחרו כ-16980 משחקים. בקבוצת הוולידציה אשר שימשה לבחירת תכונות וכיוונן פרמטרים היו \_\_\_ משחקים. בקבוצת המבחן היו .

**מספר הערות לגבי המידע:**

* המשחקים בקבוצת האימון קודמים בתאריך לקבוצת הוולידציה וקבוצת המבחן.
* השתמשנו אך ורק במשחקי ליגה, בתוצאות משחקים ב90 הדקות החוקיות(כולל זמן פציעות) מכיוון שבמשחקים מסוג זה בדרך כלל מטרת כל קבוצה היא למקסם את מספר הניצחונות ומספר השערים שלה, בעוד שבמפעלים אחרים כמו משחקי גביע לפעמים קיימת "הסתפקות במועט". למשל תוצאת תיקו בגביע אינה אומרת תיקו אלא הארכה(ואולי פנדלים) ואין סיבה לקבוצה להעדיף ניצחון על פני תיקו כזה. לעומת זאת בליגה אין לזה משמעות. יש גם משחקים עם תנאי עליה יותר מורכבים כמו בתים, תוצאות גומלין וכדומה שיכולות לגרום לכך שקבוצה לא אכפת לה להפסיד-ורצינו להימנע מהם.
* על מנת למקסם את הדמיון בין קבוצת הוולידציה לקבוצת המבחן, המשחקים לקבוצות אלה נבחרו לסירוגין כך שהמרחק בזמן של משחקי קבוצת הוולידציה וקבוצת המבחן ממשחקי קבוצת האימון פחות או יותר זהים.
* משחקי המחזור הראשון של כל עונה הושמטו מכיוון שכמעט ולא קיים מידע רלוונטי לגביהם.
* בדיעבד התברר כי בקבוצת האימון קיימת קבוצה של 302 משחקים כפולים שהוכנסו לשם בטעות. זהו מספר קטן ביחס לכלל המשחקים ולכן הנחנו שהשפעתם זניחה ונראה בזה "רעש" קטן.
* התכונות עברו נורמליזציה לפי ערכי מקסימום ומינימום של קבוצת האימון.
* אנו צופים כי משחקים במחזורים הראשונים עלולים להוות דוגמאות אימון בעייתיות מכיוון שהקבוצות עדיין לא הפגינו את היכולת שלהן בצורה מלאה. לדוגמא קבוצה חזקה שבמשחקיה הראשונים עדיין לא התבססה וביצועיה לא טובים, או שהמידע שיש כרגע עוד לא מובהק סטטיסטית. אבל מצד שני יש לה דירוג גבוה בדירוגים הכלליים שהוספנו.

עם זאת יש לציין שגם סוכנות ההימורים סובלת מאותו חיסרון בדיוק וגם עבור קובעי היחסים שלהם קיימת אותה בעיה, לכן אפשר להניח שהיתרון של סוכנות ההימורים עלינו אינו גדול יותר בשלב מוקדם של העונה.

* קבוצת הבית נוטה לנצח יחסית לקבוצות החוץ-עובדה זו נבדקה סטטיסטית על קבוצת האימון וגם על קבוצת המבחן. זו עובדה ידועה לכל חובב כדורגל. זה מתבטא בין היתר בכך שקיימים פורמטים כמו ליגת האלופות(שלא כללנו בפרויקט) בהם שערי חוץ שווים יותר מאשר שערי בית - כלומר גם התאחדות הכדורגל האירופית מודעת להטיה זו. היחס הוא: \****להשלים***\*

הסיכוי לתוצאת תיקו ולניצחון חוץ דומה.

\*יש לציין שבעונת 2020 בגלל מגפת הקורונה נוצר מצב שמשחקים אינם ללא קהל ולכן החשיבות של משחקי בית וחוץ מעט יורדת. עם זאת, מצב זה יפגע באופן דומה גם בסוכנות ההימורים וזו רק עונה אחת מבין מספר עונות שבדקנו. עדיין ראינו לנכון לציין זאת.

* במהלך כיוונון הפרמטרים השתמשנו ביחסי ההימורים של . בשלב המבחן בדקנו את ביצועי המסווג גם עבור יחסים אחרים.
* **הערה חשובה:** ישנו הבדל מהותי בין הסיווג של דוגמאות האימון לבין הפלט שאנו רוצים להשיג מהמסווג שלנו. בעוד שהראשון הוא סיווג לשלשה כלומר: ניצחון בית, תיקו או ניצחון חוץ, לעומתו, הפלט המצופה הוא הסתברות לכל אחת מהתוצאות. מדוע אנו רוצים דווקא את ההסתברות ולא להכריע מה תהיה תוצאות המשחק ואיך אנו משיגים הסתברות יבואר בהמשך. מצב זה כפי שיובהר בהמשך גורם לא אחת לתוצאות קשות יותר לפענוח ולקשרים מורכבים יחסית בין פעולת אלגוריתם הלמידה ודיוקו לבין מה יהיה הרווח שלנו בפועל.

**הצגת התכונות במודלים השונים**

ניסינו להציג את התכונות בשני אופנים:

1. פריסת כל התכונות של שתי הקבוצות כתכונות נפרדות של המשחק. בשיטה זו לא הוכנסו תכונות ה-"ראש בראש". לדוגמא מספר השערים בממוצע לעונה של קבוצת הבית היא תכונה אחרת ממספר השערים לעונה של קבוצת החוץ.

אנו צופים כי הצגה זו היא בעייתית עבור עצים, זאת כיוון שהאלגוריתם לא יודע לשייך את התכונות הקשורות אחת לשנייה. כאשר נדרש פיצול בצומת בעץ, הוא משתמש באחת התכונות מבלי להתחשב בתכונה האחרת המשלימה לה. מבחינתו כל תכונה עומדת בפני עצמה והוא לא יודע בהכרח להבין שיש תכונות שהקשר ביניהן הכרחי. למשל הוא עלול להסתכל על elo של קבוצת הבית, לראות שהוא גבוה ולהעריך שהיא תנצח - בהתעלמות מכך שייתכן שלקבוצה השנייה יש elo דומה או אפילו עדיף אבל היא קבוצת החוץ וזו תכונה שונה שנבדקת במקום אחר בעץ. עם זאת ייתכן שבמקרים מאוד מסוימים זה אולי עדיף - למשל אם האלגוריתם יזהה שקבוצה שמנצחת תמיד בבית היא קבוצה חזקה מאוד - גם אם הקבוצה השנייה מנצחת בחוץ בדרך כלל(אבל לא תמיד) וההפרש ביניהן כביכול קטן.

1. הצגת התכונות כהפרש בין שתי הקבוצות.

ההצגה השנייה היא הפרש בין תכונות קבוצת הבית לקבוצת החוץ והוספת תכונות ה-"ראש בראש" (שהיא גם סוג של הפרש). בהצגה זו אנו צופים שיפור עבור עצים מכיוון שכעת המסווג מקבל השוואת התכונות של קבוצת הבית לתכונות של קבוצת החוץ ויכול להעריך בקלות יחסית את יחסי הכוחות ביניהן.

לגבי הרגרסיה אנו מניחים כי השינוי לא יהיה משמעותי. את הקשר בין התכונות האלגוריתם עצמו קובע ע"י בחירת מקדמים. יתכן שהוא מוצא אפילו מקדמים מתאימים יותר מאשר פעולת חיסור פשוטה (מקדם ), ולמשל הוא יוכל בצורה כזאת למצוא אולי קשרים מורכבים יותר כמו הפרש בין ניצחונות של קבוצת הבית להפסדים של קבוצת החוץ שנעלמים בשיטת ההצגה השנייה.

**אסטרטגיית הימורים**

ראשית נסביר כיצד עובדים הימורי ספורט מהסוג שמכונה money line:

לכל אפשרות בהימור סוכנות ההימורים מגדירה יחס זכייה. אם אדם מהמר על אופציה כלשהי ואופציה זו אכן התרחשה אז הוא מקבל את כספו חזרה כפול יחס הזכייה. למשל יחס זכייה 3 אומר שמי ששם 10 שקלים יקבל 30 שקל חזרה (סה"כ 20 שקל רווח).

בדרך כלל יחסי הזכייה אמורים לשקף במידה רבה איזו קבוצה טובה יותר וככל הנראה נקבעים לפי שיטות שונות שחברות ההימורים מעריכות את הקבוצה הטובה יותר, שאנחנו מניחים שמשלבות בינה מלאכותית יחד עם הערכות מומחים אנושיים.

נזכיר כי יחסי ההימורים פועלים לרעת המהמר - סוכנות ההימורים "מקצצת" אחוזים מהיחס שהם חושבים שבאמת מוצדק על מנת להרוויח. בכדי להעריך את תוחלת ההפסד המצופה עבור סוכנות ההימורים שמולה אנחנו מתמודדים, בדקנו על ידי הימור רנדומלי אלף פעמים ועל ידי חישובים הסתברותיים פשוטים את גודל הקיצוץ. יש לציין שסוכנות ההימורים שונות לוקחות אחוזים שונים ולכן אין לצפות לאותן תוצאות בין סוכנויות שונות.

בהערה נוסיף שסוכנויות ההימורים לא בהכרח מקצצות אחוזים באופן זהה מקבוצות הבית והחוץ - ואין דרך טובה לדעת זאת, אלא רק להעריך כמה הקיצוץ הממוצע עבור שלישיית אפשרויות למשחק. לראיה נראה בהמשך שבמקרה של המשחקים וסוכנות ההימורים שלנו - נלקחים אחוזים שונים עבור קבוצות הבית והחוץ. ייתכן גם שישנם שיקולים נוספים בקביעת יחסים מעבר להסתברות טהורה - למשל פסיכולוגיים, לא נוכל להתחשב בכך אבל צריך להיות מודעים לאפשרות הזאת.

בניסוי זה נהמר תמיד רק על משחק בודד, ולא קומבינציות הימורים ושיטות מורכבות יותר שגורמות לתלויות בין משחקים ולסיבוך של הבעיה. העדפנו להתמקד במשחק בודד בלבד -בצורה שתשקף טוב את האלגוריתם הלומד שלנו שחוזה הסתברויות עבור משחק בודד.

קיימות מספר אסטרטגיות הימורים:

הימורים "עיוורים":

1. הימור רנדומלי. מכיוון שינן 3 תוצאות אפשריות אחוזי הפגיעה בהימור הנכון הוא .
2. *הימור על קבוצת הבית. ברוב המשחקים קבוצת הבית מנצחת. באופן דומה הימור על תיקו או על החוץ בתקווה שהיחסים עליהם גבוהים יותר.*
3. *הימור על הפייבוריטית. יחסי ההימורים נותנים אחוזי זכייה נמוכים לקבוצה הפייבוריטית. זה נותן לנו אינדיקציה מה הייתה הערכה של סוכני ההימורים.*
4. *הימור על האנדרדוג - בהנחה שאמנם הוא בדרך כלל מפסיד אבל היחס עליו גבוה יותר.*

*הערכה מושכלת של תוצאת המשחק:*

1. *הערכה על בסיס ידע מוקדם של המהמר. הכרת משחק הכדורגל ותנאיו. הכרת הקבוצות ויכולותיהן.*
2. *הימור על פי המלצות של מומחי כדורגל\הימורים.*

*בשתי האחרונות ההימור הולך על התוצאה הסבירה ביותר על סמך השיקולים שנלקחו בחשבון.* \***לא הבנתי מה אתה אומר פה\***

***האלגוריתם לבדיקת רווח – "אלגוריתם ההימור", שמשמש להערכת הביצועים***

*מבוסס די ישירות על החוויה של הימור אמיתי:  
לכל משחק מסתכלים על שלושת היחסים שהתקבלו מחברת ההימורים, ועל ההסתברויות שהאלגוריתם הלומד שלנו מחזיר.  
במידה וההסתברות כפול היחס עבור אחת מ-3 האפשרויות של ניצחון בית, ניצחון חוץ ותיקו - גדולה מ1, אז "נהמר": נאבד יחידת הימורים (שקל, דולר, כל דבר שתרצו) בודדת. במידה ויש שתי אפשרויות שהמכפלה גדולה מ-1 אז ניקח את הגדולה מבניהן. במידה ואין מכפלה גדולה מ1 אז לא מהמרים ולא מאבדים יחידת הימורים אחת.*

*לאחר מכן בודקים מה יצאה התוצאה. אם היא יצאה כמו שהימרנו אז מוסיפים למאזן יחידות הימורים לפי יחס הזכייה.*

*בסוף הימור על כל המשחקים בודקים עם כמה יחידות הימורים סיימנו ומחלקים במספר הפעמים שהימרנו כדי לדעת כמה הרווח היחסי באחוזים. אם המאזן הסופי חיובי אז הרווחנו, אם שלילי אז הפסדנו.*

***דוגמא:***

*בהינתן המשחק בין Napoli קבוצת בית לבין קבוצת החוץ שהתקיים ב-02.05.2016*

*ניתנו יחסי ההימורים הבאים:*

|  |  |  |
| --- | --- | --- |
| *בית* | *תיקו* | *חוץ* |
| 1.17 | 7.5 | 17 |

קל לראות כי הערכת סוכנות ההימורים היא לטובת קבוצת הבית. התוחלת היא הפסד ב-

.

נניח כי המסווג החזיר לנו את אחוזי ההסתברות הבאים:

|  |  |  |
| --- | --- | --- |
| *בית* | *תיקו* | *חוץ* |
| 0.6 | 0.3 | 0.1 |

*במקרה זה אנו נהמר על תוצאת התיקו אשר בתוחלת מקבלת יחס זכייה של . ספציפית כאן התוצאה הייתה ניצחון לקבוצת הבית ואז היינו מפסידים במקרה זה. הרעיון הוא שגם אם היינו צודקים הרווח לא היה משמעותי. בהימור לאורך זמן הרווח יתקבץ סביב התוחלת ונרוויח.*

*קבוצת הבית צריכה לקבל הסתברות של לפחות 0.85 כדי להיות רלוונטית.*

*אנו מניחים כי אחוזי פגיעה טובים אינם ערובה לרווח לאורך זמן-כדוגמה ניקח מקרה של קבוצה שמנצחת 90% מהמשחקים בעונה, והיחס עליהם הוא בממוצע 1.1. אם היינו מהמרים תמיד שקבוצה זו תנצח אז נסיים עם 99% מהכסף שהימרנו עליו(לאורך העונה)-כלומר הפסדנו! לעומת זאת אם אנחנו פוגעים בהסתברות של בלבד אבל יחס הזכייה שלנו הוא* 3.5  *הרי שנסיים את העונה עם רווח של 1.5%-עדיף מאשר אם היינו מהמרים יותר "שמרני".*

***השיטה שבחרנו, פרמטרי ההצלחה ולמה זה מהווה אתגר***

*בדרך כלל בעיות למידה מונחית מחולקות בגדול לשתי קבוצות: בעיות סיווג ובעיות רגרסיה.*

*הבעיה שלנו היא לא באף אחת משתי קבוצות אלה.*

*כדי לעשות רגרסיה אנחנו צריכים שיהיה לנו לכל משחק בקבוצת האימון סיווג* *שמנבא בדיוק מה היה הסיכוי האמיתי לכל תוצאה - ניצחון בית, חוץ ותיקו. הבעיה שאי אפשר לעשות כזה דבר! הנתון הזה לא קיים, שכן התוצאה של משחק היא סיווג טרינארי: בית\תיקו\חוץ, ולא ווקטור הסתברויות של מספרים ממשיים.*

*עם זאת - אנחנו לא רוצים שהאלגוריתם שלנו יחזיר רק סיווג כי אז נקבל תשובה טרינארית שמשקפת "מה התוצאה הכי סבירה" לדעת האלגוריתם שלנו, ושיטה זו לא בהכרח תניב רווח כפי שהסברנו.*

*בנוסף יש בעיה נוספת: מהי פונקציית ה*score *שלנו לפיה נדע איזה אלגוריתם ואילו פרמטרים ותכונות הצליחו? האופציה היחידה היא להמר על משחק ולבדוק אם הרווחנו. אבל משחק בודד אינו מספק מבחינה הסתברותית. כדי לשקף בצורה טובה יותר ביצועי הימורים החלטנו להמר על אלפי משחקים ולהחזיר את הרווח/הפסד המצטבר עבור האלגוריתם הלומד(והתכונות), ולהשתמש בזה כפרמטר להצלחה.*

*לסיכום: האלגוריתם הלומד שלנו מתאמן בתור בעיית סיווג, מחזיר ווקטור של 3 הסתברויות (למשחק בודד) בצורה דומה לרגרסיה - אבל הפרמטר להצלחה הוא מספר ממשי שמהווה הרווח (היחסי) עבור הימור על אלפי משחקים. זה מהווה אתגר - כי האלגוריתם לא לומד ישירות לפי פונקציית המטרה שלו ו"לא יודע" שהשימוש בו אינו ישיר. ייתכן שפרמטרים שמעלים את דיוק המסווג יגרמו דווקא לפגיעה ברווח הסופי בגלל נטיית יתר להמר תמיד על הקבוצה הטובה יותר.*

***אתגר נוסף - קשר מורכב בין איכות אלגוריתם הלמידה לבין הרווח***

*ניקח כדוגמה פשוטה למדי: נגיד שהייתה לנו דרך לדעת שהסיכוי ה"אמיתי" לתוצאה במשחק נתון הוא באמת . נניח שהיחס המתקבל מחברת ההימורים הוא .*

*אם ככה ברור שההימור המשתלם הוא תוצאת הבית (אינדקס 0).*

*אבל אם נאמר שהאלגוריתם שלנו לא מדויק במאה אחוז, ולכן קיבלנו תוצאה יחסית קרובה ליחס ה"אמיתי". במקרה הזה ההימור המשתלם הוא תוצאת החוץ ולכן האלגוריתם שלנו יבחר להמר על תוצאת החוץ.*

*מצד שני, התוצאה האם הרווחנו או לא הרווחנו מהימור כלשהו היא לא מספר ממשי רציף אלא דבר מאוד בדיד ובינארי: או שכן (ואז הרווחנו לפי היחס) או שלא(ואז לא הרווחנו).*

*כלומר תוצאות רציפות של האלגוריתם הלומד עוברות דיסקטריזציה לתוצאה בינארית של רווח או הפסד.*

*אחרי אלפי משחקים בתקווה ההסתברות אמורה להתאזן ככה שאלגוריתם טוב אכן יוביל לבחירת הימור טובה בכל משחק, ואף לרווח. עם זאת בגלל העניין הזה ייתכנו הבדלים גדולים מאוד בתוצאה הסופית של רווח או הפסד וגודלם אפילו אם ההבדל בדיוק המסווג/רגרסיה הוא קטן.*

*מסיבה זו גם לא נוכל להשתמש בפונקציית ה"כמה היינו מרוויחים" עבור ערך* result *ברגרסיה - שכן הפונקציה של הרווח תלויה ב****משחקים רבים*** *ולא באחד בודד, ולא ניתן להגדיר עבור משחק בודד כמה האלגוריתם הלומד הצליח או נכשל.*

*עם זאת חשוב להבהיר שכמובן שיש קשר בין דיוק האלגוריתם לרווח - חיזוי "מושלם" של ההסתברויות יאפשר לנצח את סוכנויות ההימורים, ובדרך כלל דיוק טוב יותר אמור לאפשר הערכה טובה יותר של איזה הימור משתלם ולאפשר רווח. עם זאת חשוב היה להבהיר שהבדלים קטנים בדיוק האלגוריתם יכולים להוביל להבדל גדול ברווח הסופי.*

***כיוונון פרמטרים***

*מספר הפרמטרים לכל מסווג הוא גדול. מספר האפשרויות לבחירת פרמטרים הוא אקספוננציאלי במספר הפרמטרים האפשריים. על מנת לברור פרמטרים מתאימים השתמשנו בשיטה הבאה:*

*הגרלת מספר גדול של קומבינציות של פרמטרים. מיון הקומבינציות לפי רווח. עבור התוצאות הטובות ביותר עם פרמטרים נתונים, להגדיר תווך חיפוש סביב אותה נקודה מתוך הנחה שלאיכות המסווג כפונקציה של הפרמטרים אין שיפועים חדים ולכן באזור של תוצאה טובה סביר למצוא תוצאות טובות אחרות, בתקווה יותר טובות.*

*כיוונון זה נעשה על קבוצת וולידציה בגודל כמעט זהה לקבוצת המבחן.*

**אלגוריתמי למידה**

בניסוי זה נעשה שימוש בספריית sklearn של פייתון, והאלגוריתמים שקיימים בה, עם מניפולציות קטנות עבור חלק משיטות הרגרסיה.

על מנת לחשב את אחוזי ההסתברות לכל אחת משלוש התוצאות האפשריות השתמשנו בשתי משפחות עיקריות של אלגוריתמי למידה:

**עצים:**

עצים בדרך כלל משמשים על מנת לסווג את הדוגמאות לסוג מסוים. על מנת לקבל הסתברות לכל אחד מהסוגים (תוצאות) השתמשנו בשיטות הבאות:

עץ עם מספר דוגמאות בעלים: במקרה כזה, כאשר מגיעים לעלה, בודקים כמה דוגמאות מכל סוג קיימות בעלה. מחלקים במספר הכולל בעלה ומקבלים הסתברות לכל אחת מהתוצאות. אם נאפשר לעצים להיפרש עד לעלים בודדים יתאפשר מצב שלא נקבל ווקטור הסתברויות אלא ווקטור שאחת התוצאות היא 1.0 ושתי האחרות הן 0.0 ו0.0. זהו לא מצב רצוי מבחינתנו לפי השיטה שהגדרנו. ולכן אנחנו מגבילים את מינימום הדוגמאות בעלה.

לא ביצענו גיזום מאוחר משום שגם ככה העצים גזומים למדי, וזמני הריצה די גדולים.

בכל פיצול של העץ אפשרנו שימוש בכל סט התכונות הנוכחי כדי שלא להוביל לעץ בודד ומאוד לא יציב שתלוי מאוד בבחירה אקראית. כמו כן השתמשנו במשקול - מכיוון שניצחונות בית נפוצים בהרבה מאשר שאר התוצאות הרי שיכולה להיות הטיה של התוצאות לטובת בית. משקלנו לפי היחסים בקבוצת האימון בין ניצחון בית, תיקו וניצחון חוץ.

לרוב עצים נוטים להתאמת יתר ושונות גדולה, אבל בכל זאת ננסה להשתמש גם בעצים בודדים, ובנוסף עצים טובים יכולים לשמש כבסיס טוב לgradient boost, כפי שיבואר בהמשך.

יער אקראי:

במקרה זה ניתן לתת לכל אחד מהעצים להתפרש במלואם, אם כי לא חובה לאפשר זאת כי זה יכול לגרום להתאמת יתר (בדקנו פרמטרים על עומק מקסימלי לעץ ומספר התכונות בעלה), ואחוזי העלים של כל העצים ביחד הם ההסתברות המבוקשת. השתמשנו באלגוריתם כפי שהוא מופיע בויקיפדיה, כלומר בכל שלב באלגוריתם הפיצול הוא לפי התכונה הטובה יותר מבין תת קבוצה אקראית בגודל שורש של מספר התכונות או בגודל של מספר התכונות. בנוסף נעשה שימוש בשיטת ה-bootstrap עם max samples(כלומר כל עץ מאומן על מספר משחקים זהה למספר הכולל, אבל עם חזרות על חלק מהמשחקים באקראי כמה פעמים).

גם פה נעשה שימוש במשקול דוגמאות האימון.

המטרה היא שכל עץ יפצה על הטעויות של עצים אחרים ובאמצעות מספר רב של עצים נוכל להגיע לאלגוריתם יציב, עם שונות קטנה יותר אבל עם הטיה נמוכה.

Extra random forest:

אלגוריתם דומה מאוד לקודם, עם שני הבדלים:

* לא עושים boostrap אלא לוקחים את כל המשחקים עבור כל עץ.
* נקודת הפיצול עבור כל תכונה היא אקראית ולא נקודת הפיצול האופטימלית. עם זאת עדיין נבחרת תמיד התכונה הטובה ביותר על פי מדדי gini/entropy מבין התכונות הנבדקות עבור הפיצול שנבחר.

שינויים אלה אמורים לפי הדוקומנטציה של sklearn להקטין מעט את השונות ולהגדיל מעט את ההטיה ביחס לעצים אקראיים רגילים.

Gradient boosting:

נעשה שימוש באלגוריתם זה, על בסיס 3 מסווגים שונים:

1. כלום - מסווג האפס.
2. מסווג טיפש – ברירת המחדל. מסווג נאיבי שלא משתמש בשום אלגוריתם.
3. עץ שלנו - עבור כל מספר וסוג תכונות נבחר העץ הבודד שנתן את התוצאה הטובה ביותר עבור תכונות אלה.

יש לציין שעצי הרגרסיה של האלגוריתם Gradient bossting משתפרים בכיוון הגרדיאנט, בהנחה של פונקציית loss של **מסווג**, שזו אינה פונקציית ההפסד/רווח שלנו. כלומר, ייתכן שהעצים שיצאו יהיו מאוד מדויקים מבחינת בחירת המנצח אבל יתנו לו סיכוי טוב מידי, וסיכוי נמוך מידי לקבוצה הפחות טובה לדעתם. לכן אנחנו מניחים שאלגוריתם זה יבחר באחוזים גבוהים יותר בקבוצה הטובה ובהתאם לכך יזכה באחוזים גבוהים יותר, אם כי לא בהכרח ביותר כסף - הימור מאוד "שמרני".

בסופו של דבר האלגוריתם מחזיר את ההסתברות המשוקללת לפי יער עצי הרגרסיה שנוצרו, עם משקל שונה לכל אחד מהעצים בהתאם למדד learning rate שנקבע מראש.

לא בדקנו את כל הפרמטרים הרבים של אלגוריתם זה, אבל כן בדקנו פרמטרים שעוסקים ביער העצים שנוצרו (מספר עלים, עומק וכו'). כמו כן, ניסינו מספר ערכים של learning rate שקטנים או שווים מ0.1 לפי המלצה של הדוקומנטציה, וגודל שונה של דגימות שישמשו ללמידת כל אחד מהעצים – סוג של וולידציה שמקטינה את השונות.

מטעמי זמני ריצה גדולים בחרנו שמספר התכונות בהן ישתמש הלאגוריתם ביצירת פיצול בעץ הוא ביחס לסך התכונות, ולא ערכים גדולים יותר, שכן על פי המלצת הספרייה sklearn זה מזרז את זמני הריצה.

**רגרסיה:**

שיטה זו משתמשת בדרך כלל לניבויים וכמו כן להערכת הסתברויות. השתמשנו בשיטות הבאות:

רגרסיה לוגיסטית: שיטה זו מניחה קשר לינארי בין התכונות לבין הסיווג. על בסיס דוגמאות האימון היא מחפשת מקדמים מתאימים. שיטה זו מיועדת בעיקרה לסיווג בינארי. על מנת להשיג סיווג טרינארי השתמשנו בשיטות הבאות:

1. רגרסיה מולטי – הרחבה של שיטת הרגרסיה הלוגיסטית ליותר משני סוגים.
2. One vs all: בשיטה זו בונים שני מסווגים: האחד מחשב את סיכויי ההצלחה של קבוצת הבית, השני מחשב את סיכויי ההצלחה של קבוצת החוץ ומה שנשאר הוא הסיכוי לתוצאת תיקו.

בשיטה זו מתקבלת משוואה לינארית הקושרת בין ערכי התכונות (משתנים מסבירים) לבין הערך הנלמד (משתנה מוסבר). נקבע ערך סף עבור המשתנה המוסבר. ע"י הצבה של ערכי תכונות של דוגמא שאותה רוצים להעריך והשוואה לערך הסף

\****תרחיב יותר כיצד מחושבות ההסתברויות אולי? נשקול את זה***

ובכל אחת מהשיטות בדקנו מספר פותרים שונים עם פרמטרים שונים. כמו כן בדקנו עם התוצאות עם ובלי משקול לתוצאות - שכן בניגוד לעצים זמני הריצה פה לא גדולים כל כך ויכולנו להרשות לעצמנו לבדוק יותר אפשרויות פרמטרים.

רגרסיה לינארית: באופן דומה לרגרסיה הלוגיסטית, גם כאן מניחים קשר לינארי בין התכונות לסיווג. בשונה מהלוגיסטית, תוצאת המסווג היא מספר. על מנת לחשב הסתברות, לקחנו את התוצאה וחישבנו את המרחק היחסי של המספר לכל אחד מהסיווגים של דוגמאות האימון . מרחק זה חלקי סכום המרחקים נותן לנו הסתברות.

ברגרסיה לינארית נבדקו כמה solvers שונים, וכמה שיטות שונות לרגורליזציה ועונש. יש לציין שאמנם המימוש של הרגרסיה הלינארית עצמה הוא של sklearn אבל ההמרה להסתברויות היא אלגוריתם שלנו.

***בחירת תכונות:***

*השתמשנו בשיטת* RFE-Recursive Feature Eliniation. *בשיטה זאת עבור* k *שלם בכל איטרציה נבדק ה*gini *במקרה של עץ או המקדמים במקרה של רגרסיה, ומסוננת התכונה עם הדירוג הנמוך ביותר, וממשיכים רקורסיבית עד למצב של הגעה למספר התכונות הנדרש. בצורה זו אמורים להגיע לקבוצת תכונות שתיתן תוצאה טובה יותר מאשר סתם לקחת את k המובילים מבחינת* gini/coef.

*החלטנו לבדוק עבור 100% תכונות, 90%, 75%, 50%, 25%.*

*כחובבי כדורגל והמשחק* fifa *עניין אותנו מה יהיו התכונות המובילות בכל שיטת הצגה, ואיפה נמצא דירוג הקבוצה במשחק.*

*נסתכל בנפרד על כל שיטת הצגה ועל רגרסיות ועצים.*

*שיטת ההצגה הראשונה:*

*עבור עצים, שלא במפתיע התכונות המשמעותיות ביותר בפער מבחינת* gini *הן ה*elo *של קבוצת הבית והחוץ.* *זה היה צפוי בגלל שמספר זה מתחשב בנתונים רבים וידוע כפרמטר חזק לקביעת עוצמת קבוצות.*

*תכונות אחרות שיחסית משמעותיות אמנם בפער פחות מ*elo *, וקרובות זו לזו:*

*תכונות שקשורות למספר בעיטות, בעיטות למסגרת ושערים והפרשים שלהם – שערים לטובת ונגד, בעיטות לקבוצה ונגדה...*

*רוב התכונות מהסגנון הזה בעלות* gini *דומה מאוד זו לזו והסבר אפשרי לכך הוא שרוב התכונות דומות או חופפות מבחינת המידע.*

*פרמטרים דומים עבור קבוצת הבית והחוץ לא תמיד קרובים אחד לשני אם כי יש מקרים שכן, מה שגורם למשל שבחלק מקבוצות התכונות ייתכן שיישאר נתון מסוים עבור קבוצת הבית ולא עבור קבוצת החוץ מה שיכול להוביל לדיוק רע של העצים. למשל עבור 75% התכונות ציוני המשחק FIFA עבור הקבוצה הביתית כלולים אבל לא עבור קבוצת החוץ. עם זאת לא נראית קורולציה מובהקת האם יש חשיבות גדולה יותר לפרמטרים של קבוצת חוץ או הבית.*

*לאכזבתנו נראה שציוני FIFA אינם פרמטר מאוד חזק מבחינת* gini *ובסינון ל50% מהתכונות הם מסוננים. במפתיע גם תכונות של רצף משחקים אחרונות ללא הפסד/ללא ניצחון, והצלחה ב5 המשחקים האחרונים התבררו כפרמטרים לא מאוד משמעותיים בניגוד לציפיות שזה ישקף את הכושר הנוכחי של הקבוצות.*

*תכונה מפתיעה מבחינתנו היא הפרש הכרטיסים האדומים עבור קבוצת הבית - נראה שאם קבוצה נוטה לכרטיסים אדומים רבים נגדה, ומשחקת בבית אז זה מגדיל את סיכוייה לנצח.*

*במידה מסוימת עם זאת התוצאות צפויות – קשה להעריך כמה מוסיף מידע על קבוצה בודדת בלי מידע קשור על הקבוצה היריבה, ועקב כך רוב התכונות קשה להעריך את תוספת המידע שלהן לבדן לבניית עץ.*

*עבור רגרסיות התוצאה קצת שונה: אחוז הניצחונות, ההפסדים, התיקו והנקודות הם הפרמטרים המשמעותיים ביותר מבחינת משקל בתוצאה הסופית. כאשר מסננים תכונות אז ב25% וב50% נראית גם חשיבות גדולה לחלק מהתכונות של תוצאות 5 המשחקים האחרונים מה שהיה מאוד לא משמעותי בעצים. עם זאת גם ברגרסיות החשיבות של רצף ניצחונות/הפסדים וכדומה אינו גבוה.*

*לעומת זאת תכונות ה*elo *לא כל כך משמעותי, וגם כאן ציוני המשחק FIFA מקבלים חשיבות בינונית בלבד.*

*מעניין לציין שתכונות רבות שקשורות לכרטיסים אדומים הגיעו ל25% המובילות...*

*בניגוד לעצים כאן לרוב תכונות "דומות" (אותה תכונה) או מנוגדות (ניצחונות של קבוצת הבית מול הפסדים של קבוצת החוץ) בין קבוצות הבית והחוץ לרוב מקבלות חשיבות דומה מאוד – וזה די סביר, בניגוד לעצים כאן האלגוריתם שמבוסס על שיטות של אלגברה לינארית יודע להבין טוב יותר קשרים בסגנון הזה.*

*שיטת ההצגה השנייה:*

***בחירת שילוב הפרמטרים והתכונות עבור כל אלגוריתם להרצה על קבוצת המבחן***

*ניסינו על קבוצת הוולידציה, עבור כל אלגוריתם סט תכונות מסוים (על פי האחוזים לעיל) - שילובים שונים של פרמטרים.*

*כפי שצוין לעיל לא לקחנו משחק בודד אלא בדקנו מה יוצא הרווח או הפסד המצטבר עבור כל קבוצת הוולידציה בתור הפרמטר להצלחת האלגוריתם.*

*במקרה של רגרסיה ועצים בודדים ניסינו את כל שילובי הפרמטרים האפשריים (בפרמטרים ששינוי שלהם יכול להשפיע על דיוק האלגוריתם) כדי לבחור את התוצאות הטובות ביותר.*

*ביערות ו*gradient boost  *בגלל זמני הריצה הגדולים יצרנו שילובי פרמטרים מתוך המכפלה הקרטזית של הפרמטרים האפשריים, מספר רב של פעמים - לעיתים 500 קומבינציות ולעיתים 1000 תלוי בזמן הריצה. אנו מניחים שדגימה אקראית זו תתפרס בצורה כזו שניתן יהיה להעריך את איכות האלגוריתם וניתן יהיה מתוך כל כך הרבה קומבינציות פרמטרים למצוא שילוב טוב ולהסיק מסקנות.*

*עבור כל אלגוריתם לקחנו בחשבון את השיקולים הבאים*

* *מה שילוב הפרמטרים ותכונות שנותן את התוצאה הטובה ביותר - בתקווה ששילוב זה ייתן תוצאות טובות גם בקבוצת המבחן.*

*עם זאת, כדי למנוע מצב שאנחנו מכנים "התאמת יתר פרמטרית" ובו שילוב פרמטרים מסוים מצליח מאוד עבור קבוצת הוולידציה אבל לא בהכרח לקבוצת הטסט - כלומר שונות גדולה מידי ואלגוריתם לא יציב, התחשבנו גם עבור "קבוצת אלגוריתמים דומים"(למשל סט תכונות שונה, או שיטות שונות של* gradient boost)*:*

* *ממוצע גבוה עבור כל הקומבינציות שנוסו - בהנחה שאם אלגוריתם יראה ביצועים טובים עבור מספר רב של שילובי פרמטרים אז גם עבור קלט שונה הוא יצליח טוב שכן זה מראה שלא רק שילוב פרמטרים ספציפי מצליח עבורו אלא שהאלגוריתם עצמו מתאים למשימה.*
* *שונות נמוכה של הרווח: בהנחה שאלגוריתם שיוכיח "יציבות" עבור שילובים שונים גם יראה יציבות עבור קלט שונה - קבוצת המבחן, ולא תלוי מאוד בקלט הספציפי מה שייתכן שמראה על אלגוריתם לא יציב ולא אמין מספיק גם אם יש לו לפעמים תוצאות טובות.*

*שתי הנחות אלה כמובן לא בהכרח נכונות, אבל לקחנו אותן בחשבון.*

* *במידה ויש מספר תוצאות שסיפקו רווח טוב, אבל פרמטרים שונים מאוד אחד מהשני עד כדי פעולת אלגוריתם שונה לחלוטין - לפעמים נלקחו שתי הקומבינציות הטובות. במקרה כזה לרוב נעדיף לקחת שילובי פרמטרים עבור מספר תכונות* ***שונה***
* *במידה והשונות ברווח נמוכה יחסית - אבל נראה שהקומבינציה המוצלחת ביותר הרוויחה באופן חריג יחסית לשאר הקומבינציות שנוסו: אז לעיתים נבדוק גם את "המקום השני" או עבור סט תכונות שונה.*

*פירוט מלא של האלגוריתמים עבור כל אחד מהאלגוריתמים שנבחרו מצורף בנספחים. כמו כן, האלגוריתמים נתונים להרצה בקובץ: first\_method\_algo.py*

***תוצאות הניסוי ודיון***

*אלגוריתמים נאיביים:*

*לפני שנשווה את ביצועי האלגוריתמים החכמים יותר אנחנו נשווה שיטות הימורים נאיביות -כמה הן הרוויחו או הפסידו.*

*הפרמטר העיקרי להצלחה הוא הרווח אבל יש עניין גם בבדיקה אחוז הפגיעה בהימור הנכון ביחס לסך המשחקים שהימרנו. לכן בכל הבדיקות בטסט נכלול גם את המספרים האלה. ברצוננו גם לבחון האם אלגוריתם שמהמר נכון באחוזים גבוהים בהכרח יגיע לרווח או לא.*

*תוצאות:*

|  |  |  |
| --- | --- | --- |
| *שם* | *רווח (%)* | *אחוז פגיעה(%)* |
| *הימור רנדומלי (ממוצע על 1000 הרצות)* |  |  |
| *הימור על התוצאה הפחות סבירה* |  |  |
| *הימור על תוצאת הביניים* |  |  |
| *הימור על התוצאה העדיפה* |  |  |
| *הימור על קבוצת הבית* |  |  |
| *הימור על תוצאת תיקו* |  |  |
| *הימור על קבוצת החוץ* |  |  |

מסקנות ביניים לגבי קבוצת המבחן:

בקבוצת המבחן נראה יתרון ברור למי שמהמר תמיד על הקבוצה הביתית - עד כדי רווח. זאת למרות שאחוז הניצחונות עבור הקבוצה הביתית שהוא 46.67% לא שונה משמעותית ממה שהיה בקבוצת האימון (46.32%) ולמרות ששיטת הימור דומה על קבוצת הוולידציה שגם בה האחוזים דומים (45%) לא הניבה רווח אלא הפסד של 4.4%.

זה שאומר שהסיבה להבדל היא שמשום מה למרות שבחירת המשחקים לוולידציה ולטסט הייתה לסירוגין לפי אנדקסי משחקים הרי שנוצר מצב מוזר שדווקא בקבוצת הטסט היחסים עדיפים משמעותית על קבוצת הבית ביחס לקבוצת הוולידציה - שלפיה עשינו כוונון פרמטרים ותכונות.

כמובן שאם היחסים על הבית מאוד טובים אז כדי שחברת ההימורים תרוויח היא צריכה לאזן את זה על ידי יחסים נמוכים יחסית עבור שאר האפשרויות. כלומר הימור על תיקו או ניצחון חוץ בקבוצת הוולידציה עדיף בעיקרון מאשר לעשות הימור דומה בקבוצת המבחן.

לכן אנחנו חוששים שיהיה לנו מצב של הטיה והתאמת יתר פרמטרי/תכונות שיגרום לכך שהצלחת האלגוריתמים בקבוצת הוולידציה לא תוביל בהכרח להצלחה בקבוצת המבחן, ולהפסדים משמעותיים. עם זאת במידה ואלגוריתם שלנו ידייק בחיזוי עד כדי כך שהוא יצליח למרות הטיה זו – הרי שזה מבסס משמעותית את יציבות האלגוריתם.

\*גם בקבוצת המבחן וגם בקבוצת הוולידציה הימור על הבית נותן תוצאות עדיפות, אבל בקבוצת המבחן הפער גדול בהרבה עד כדי רווח קטן בהימור על הבית.

בנוסף נשים לב שהימור על הפייבוריט עדיף מאשר הימור על האנדרדוג. זו תוצאה צפויה משום שלרוב קבוצת הבית היא הפייבוריטית. כמו כן נשים לב שלמרות שהימור על הפייבוריט מנצח באחוזים טובים יותר של 55.42% עדיין הרווח עדיף בהימור על הבית מה שמראה שלא בהכרח להמר על הקבוצה שנראית עדיפה היא הדבר הנכון - כפי שצפינו.

המדד להצלחה עבור האלגוריתמים שנבחרו:

* ניצחון על מהמר אקראי – מוכיח כי ידע מוקדם עוזר למהמר
* ניצחון על כל השיטות הנאיביות – מוכיח כי ידע מוקדם עדיף על שיטות הימור פשוטות (אף אם הגיוניות)
* רווח גבוה ככל הניתן בסיכום הכללי עבור כל המשחקים
* רווח לאורך זמן – התבססות מהירה סביב ערך התוחלת – כמה זמן עד שהרווח הוא קבוע.

*תוצאות עבור האלגוריתמים שנבחרו בשיטת הצגת תכונות ראשונה:*

*סך התכונות בשיטה הראשונה הוא 102.*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| *אלגוריתם* | *מספר תכונות(%)* | *אחוז פגיעה במבחן (%)* | *רווח על קבוצת המבחן (%)* | *רווח על קבוצת הוולידציה (%)* |
| *עץ* |  |  |  |  |
| *עץ* |  |  |  |  |
| *יער* |  |  |  |  |
| *יער* |  |  |  |  |
| *יער* |  |  |  |  |
|  |  |  |  |  |
| *גרדיאנט- zero* |  |  |  |  |
| *גרדיאנט- none* |  |  |  |  |
| *גרדיאנט- tree* |  |  |  |  |
| *גרדיאנט- tree(1)* |  |  |  |  |
| *גרדיאנט- tree(2)\** |  |  |  |  |
| *רג' ליאנרית Ridge* |  |  |  |  |
| *רג' ליאנרית Elastic* |  |  |  |  |
| *רג' לוגסטית- מולטי-saga* |  |  |  |  |
| *רג' לוגסטית- מולטי-saga* |  |  |  |  |
| *לוגיסטית- one vs all* |  |  |  |  |

*\* קומבינציית פרמטרים שונה, במקרה של הבדל משמעותי בפרמטרים*

*המסקנה משתי הטבלאות מאד ברורה: כישלון גדול בכל מדד שהגדרנו. התוצאות של קבוצת הוולידציה וקבוצת המבחן שונות כמעט לחלוטין ומרגישות כמעט אקראיות. נראה כי האלגוריתמים בשיטת ההצגה הראשונה מאוד לא יציבים ותלויים מאוד בפרמטרים הספציפיים שיתאימו עבור סט משחקים ספציפי – התאמת יתר פרמטרית וייתכן שגם התאמת יתר במידע.*

*ישנו שיפור קל ביחס למהמר אקראי באלגוריתמים מבוססי גרדיאנט, רגרסיה לינארית ומולטי אך לא בכולם, לכן סביר להניח שזו הצלחה מקרית.*

*נשים לב שאמנם גם הפעם לא בהכרח אחוזי ניצחון טובים אומרים רווח עדיף אבל כן יש מגמה מסוימת כזאת - בהתחשב בכך שאם פוגעים באחוזים גבוהים אז זה לרוב אומר שהימרו על הפייבוריט שראינו שהיחסים עליהם עדיפים באתר 365bet בקבוצת המבחן.*

*נראה שלא סתם חברות ההימורים ממשיכות להרוויח. לנצח אותם זו משימה קשה. אולי המאמר שהזכרנו בתחילת הדוח צדק ואין חשיבות רבה לידע מוקדם בהימורי ספורט?*

*אבל נשארה לנו שיטת ההצגה השנייה. שאנחנו מעריכים שתצליח טוב משמעותית. האם בעזרתה נוכל לנצח את חברות ההימורים, למרות התפלגות יחסים שונה ממה שציפינו? או שמא הדבר הכי טוב לעשות הוא באמת להמר באופן נאיבי?*

*תוצאות עבור האלגוריתמים שנבחרו בשיטת הצגת תכונות שנייה:*

*סך התכונות בשיטה השנייה הוא 54.*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| *אלגוריתם* | *מספר תכונות(%)* | *אחוז פגיעה במבחן (%)* | *רווח על קבוצת המבחן (%)* | *רווח על קבוצת הוולידציה (%)* |
| *עץ* |  |  |  |  |
| *עץ* |  |  |  |  |
| *יער* |  |  |  |  |
|  |  |  |  |  |
| *גרדיאנט- zero* |  |  |  |  |
| *גרדיאנט- none* |  |  |  |  |
| *גרדיאנט- tree* |  |  |  |  |
| *רג' ליאנרית Ridge* |  |  |  |  |
| *רג' ליאנרית Elastic(1)* |  |  |  |  |
| *רג' ליאנרית Elastic(2)* |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |

*ניתן לראות באופן מובהק ששיטה זו מוצלחת יותר.*

*ביחס בין אחוזי פגיעה לרווח: ניקח לדוגמא את התוצאה עבור יער ו-extra יער: הם פוגעים באחוזים דומים מאד, אבל התוצאה הסופית שונה בכמעט . נשים לב שעבור שניהם אחוזי הפגיעה פחות טובים מאשר הימור אקראי בין שלוש אפשרויות אשר בממוצע מתכנס סביב ובכל זאת שניהם מצליחים לעקוף מהמר אקראי. מכאן ניתן להסיק כי לא די באחוזי פגיעה על מנת להרוויח מחד, ומצד שני אחוזי הצלחה נמוכים, לא מעידים בהכרח על הפסד.*

*על מנת לבדוק האם זו הצלחה מקרית, או הצלחה לאורך זמן, חילצנו את הגרפים הבאים המציגים את הרווח כפונקציה של מספר ההימורים שבוצעו. במקרה זה הצגנו גרפים רק עבור שלושת האלגוריתמים המובילים:*

**

*ניתן לראות כי החל משלב מסוים הרווח מתכנס לכיוון התוחלת. מכיוון שהתוחלת חיובית, מובטח מצב של רווח לאורך תקופה. כמו כן ניתן להבחין כי הגרפים מאופיינים בעיקר בעליה קיצונית בהתחלה ולאחר מכן ירידה חדה, התנדנדות ואז התכנסות איטית לכיוון הערך הסופי. אם נזכור את אחוזי הפגיעה שהתקבלו, הרי שברוב ההימורים אנחנו מפסידים, אבל המגמה הכללית, כפי שניתן לראות בגרף שומרת על רווח חיובי.*

*על מנת להבין את סדר הגודל של הזמן מבחינת מספר הימורים: בכל אחת מחמש הליגות הבכירות שהתייחסנו אליהן יש כ-370 משחקים בעונה (כאשר הורדנו את המשחקים בסיבוב הראשון). כלומר בשנה אחת ניתן להמר על כ-1850 משחקים. (תחת הנחה שניתן להמר על כל משחק בפני עצמו).*

***מסקנות***

*מהניסוי הגענו לכמה מסקנות מעניינות:*

* *שיטת ההצגת התכונות הראשונה מובילה להפסד*
* *שיטת הצגת התכונות השנייה מובילה בד"כ לרווח עבור עצים*
* *השיטה עובדת עבור סוכנויות שונות*
* *הקשר בין רווח לאחוזי פגיעה, תלוי מאד ביחס ההימורים שניתנו. אחוזי פגיעה טובים לא מבטיחים רווח, ואילו אחוזי פגיעה נמוכים לא מעידים על הפסד. יחד עם זאת קיים קשר בין השניים.*
* *שיטת עצים עדיפה על רגרסיה: תוצאה זו מקבילה לתוצאה שהתקבלה במאמר שהזכרנו.*

***סיכום ורעיונות להמשך מחקר***

*הימורים על משחקי ספורט הם תופעה קיימת וידועה. אנשים רבים מכלים את כספם על הימורים מסוג זה, מתוך הנחה שהם יכולים להרוויח על בסיס בקיאותם בתחום. מחקר שהבאנו[1] טוען שזו הנחה שגויה מכיוון שאחוזי הפגיעה שלהם לא טובים משל אנשים שאינם מכירים את התחום.*

*בניסוי זה הוכחנו באופן אמפירי שניתן להרוויח לאורך זמן על בסיס ידע מוקדם, ע"י סידור נכון של ידע זה, המשווה באופן מלא בין שתי הקבוצות ומגדיר את היחס בין היכולות שלהן. לעומת זאת ראינו כי ידע שלא מסודר כראוי ולא עורך השוואה נכונה בין הקבוצות עלול להוביל להפסד גדול אף ביחס למהמר אקראי. כמו כן ראינו שאחוזי פגיעה אינם מדד בלעדי להצלחה.*

*הרווח הגדול ביותר שהשגנו עבור קבוצת המבחן עומד על כ-20%. רווח זה גדול יותר פי 12 מהרווח שראינו עבור שיטה המהמרת על בסיס השוואה בין שחקנים[2].*

*התוצאות מעודדות אך יש עוד לא מעט כיוונים להמשך מחקר:*

*בחירת תכונות:*

*על פי מאמר של אלטמן ושות'[6] לשיטת הדירוג התכונות הרגילה, יש נטייה להעדיף תכונות אשר מספר הערכים השונים שלהן הוא גבוה מאד. הם מציעים את שיטת הפרמוטציות: בשיטה זו מבצעים את אותו דירוג אבל כאשר אחת התכונות "מבולגנת", כלומר נבחרת פרמוטציה שונה שלה ביחס לסיווג. ע"י השוואה בין דירוג התכונה באופן רגיל לדירוג התכונה בשיטה זו, ניתן להעריך עד כמה הקשר בין התכונה לסיווג הוא חזק.*

*לדוגמא במקרה שלנו לרוב התכונות יש מספר ערכים שונים הנע בין כמה אחדות לכמה אלפים. לעומת זאת לתכונת ה-elo יש ערך שונה כמעט עבור כל דוגמא (כ-16000). על פי שיטה זו, ניתן לבדוק עד כמה איכות תכונת ה-elo באמת חשובה.*

*דרך נוספת להורדת מספר התכונות היא למצוא את רמת הקשר בניהן ע"י* מקדם המתאם של ספירמן. תכונות התלויות אחת בשנייה ברמה גבוהה יכולות להיות מצומצמות לתכונה מייצגת אחת.

*יצירת תלויות:*

*סוכנויות הימורים רבות (למשל בארץ), לא מאפשרות להמר על משחק בודד על מנת להקשות על מהמרים להרוויח. כיצד משפיעות תלויות כאלה על האלגוריתם שלנו? האם יש דרך אחרת להתמודד עם מתכונת הימורים כזאת?*

*למידה על תהליך ההימור:*

*בניסוי זה אסטרטגיית ההימור היא שלנו, על בסיס הסתברויות של האלגוריתם. האם ניתן ללמוד אסטרטגיית הימור טובה? מה המדדים לבחירת אסטרטגיה כזו?*

***ביבליוגרפיה***

# *[1]* Football gambling three arm-controlled study: gamblers, amateurs and laypersons, [Ronen Huberfeld](https://pubmed.ncbi.nlm.nih.gov/?term=Huberfeld+R&cauthor_id=22890307), [Roman Gersner](https://pubmed.ncbi.nlm.nih.gov/?term=Gersner+R&cauthor_id=22890307), [Oded Rosenberg](https://pubmed.ncbi.nlm.nih.gov/?term=Rosenberg+O&cauthor_id=22890307), [Moshe Kotler](https://pubmed.ncbi.nlm.nih.gov/?term=Kotler+M&cauthor_id=22890307), [Pinhas N Dannon](https://pubmed.ncbi.nlm.nih.gov/?term=Dannon+PN&cauthor_id=22890307), Psychopathology, August 12

# [2] Machine Learning in Football Betting: Prediction of Match Results Based on Player Characteristics Johannes Stübinger 1,\* , Benedikt Mangold 2 and Julian Knoll

[3]<https://www.football-data.co.uk/>.

[4] http://clubelo.com/

[5] האתר של ה fifa

# [6] Permutation importance: a corrected feature importance measure ,[André Altmann](javascript:;), [Laura Toloşi](javascript:;), [Oliver Sander](javascript:;), [Thomas Lengauer](javascript:;) , [Author Notes](javascript:;) Bioinformatics, Volume 26, Issue 10, 15 May 2010, Pages 1340–1347

***נספחים***

*רשימת תכונות והסבר למה הן (אולי) משמעותיות:*

*הערה – במ"ל – בממוצע למשחק.*

*במל"ב – בממוצע למשחק בית.*

* *דירוג elo לקבוצת הבית: הוסבר לעיל.*
* *דירוג fifa לקבוצת הבית: הוסבר לעיל.*

*לכל התכונות הבאות יש תכונה דומה נוספת המתייחסת רק למשחקי הבית עבור קבוצת הבית. בנוסף לכל התכונות יש תכונה מקבילה עבור קבוצת החוץ ותכונה דומה עבור משחקי החוץ של קבוצת החוץ. הסיבה היא שייתכן שיש הבדל בהתנהגות של קבוצה בבית ובחוץ מבחינת נטייה לתוצאות מסוימות וכדאי להתייחס לזה בנפרד מהביצועים של הקבוצה במשחקי הבית והחוץ יחדיו.*

* *מספר ניצחונות לקבוצת הבית במ"ל: ככל שקבוצה נוטה יותר לנצח ככה סיכוייה לנצח משחק הבא כנראה גדולים יותר.*
* *מספר תוצאות תיקו לקבוצת הבית במ"ל: ככל שקבוצה נוטה יותר לתוצאות תיקו ככה סיכוייה לתיקו כנראה גדולים יותר.*
* *מספר הפסדים לקבוצת הבית במ"ל: ככל שקבוצה נוטה יותר לתוצאות תיקו ככה סיכוייה להפסידו כנראה גדולים יותר.*
* *מספר שערים לקבוצת הבית במ"ל: קבוצה שמבקיעה הרבה הסיכוי שלה לנצח לרוב גדול יותר והסיכוי להפסד קטן יותר-הרי מי שמבקיע יותר מנצח בכדורגל.*
* *מספר שערים שספגה קבוצת הבית במ"ל: קבוצה שסופגת הרבה הסיכוי שלה להפסיד לרוב גדול יותר והסיכוי לנצח קטן יותר-הרי מי שסופג יותר מפסיד בכדורגל*
* *הפרש שערים לקבוצת הבית: קבוצה עם הפרש שערים חיובי נוטה להבקיע יותר מאשר שהיא סופגת ולכן סיכוייה לנצח גדולים יותר, וסיכוייה להפסיד קטנים יותר.*
* *מספר נקודות למשחק לקבוצת הבית במ"ל: מספר הנקודות הוא פונקציה של מספר הניצחונות, התיקו וההפסד ומהווה המדד לאיזו קבוצה זכתה באליפות בסוף העונה-אז אנחנו מניחים שהוא גם מדד מוצלח לרמת הקבוצה וסיכוייה לנצח.*
* *מספר משחקים עם רשת נקיה לקבוצת הבית: קבוצה ששמרה על רשת נקיה בהכרח לא יכולה להפסיד. אז קבוצה עם נטיה לשמור על רשת נקיה כנראה שלא תפסיד.*
* *מספר משחקים בלי הבקעות לקבוצת הבית: קבוצה שלא הבקיעה בהכרח לא יכולה לנצח באותו משחק, אז קבוצה שנוטה שלא להבקיע כנראה שלא תנצח.*
* *מספר בעיטות במ"ל: ובדרך כלל מספר בעיטות גבוה מתורגם למספר שערים גבוה ולאחוז ניצחונות גבוה יותר. מכיוון שקיים בכדורגל אלמנט כלשהו של "מזל" מבחינת מספר השערים שהובקעו בפועל, הרי שמספר הבעיטות לפעמים משקף טוב יותר את איכות ההתקפה של הקבוצה מאשר מספר השערים בפועל.*
* *מספר בעיטות למסגרת במ"ל של קבוצת הבית: בדומה לקודם, רק שבעיטות למסגרת הינן מדד טוב עוד יותר לרמת ההתקפה של הקבוצה ביחס ליריבותיה, שכן בעיטות למסגרת יתורגמו לשער אלא אם השוער של הקבוצה היריבה עצר את הבעיטה.*
* *מספר בעיטות של היריבה במ"ל, עבור קבוצת הבית: בדומה לקודם, משקף את רמת ההגנה של הקבוצה.*
* *הפרש בין הבעיטות במשחקיה של קבוצת הבית, לטובת הקבוצה, לטובת הבעיטות של היריבה במשחקיה של קבוצת הבית, במ"ל: מראה האם הקבוצה לרוב מאיימת על השער יותר מאשר שהיא מאוימת, ולכן האם היא עשויה להבקיע יותר שערים במשחק מאשר שהיריבה תבקיע (ולכן תנצח).*
* *הפרש בין הבעיטות במשחקיה של קבוצת הבית, לטובת הקבוצה, לטובת הבעיטות של היריבה במשחקיה של קבוצת הבית, במ"ל: מראה האם הקבוצה לרוב מאיימת על השער יותר מאשר שהיא מאוימת, ולכן האם היא עשויה להבקיע יותר שערים במשחק מאשר שהיריבה תבקיע (ולכן תנצח).*
* *הפרש בין מספר הבעיטות למסגרת במשחקיה של קבוצת הבית, לטובת הקבוצה, לבין מספר הבעיטות למסגרת של היריבה במשחקיה של קבוצת הבית, במ"ל: מראה האם הקבוצה לרוב מאיימת על השער יותר מאשר שהיא מאוימת, ולכן האם היא עשויה להבקיע יותר שערים במשחק מאשר שהיריבה תבקיע(ולכן תנצח).*
* *כרטיסים אדומים לקבוצת הבית במ"ל: קבוצה שנוטה לקבל כרטיסים אדומים נוטה להיכנס לחיסרון מספרי מה שפוגע בסיכוייה לנצח, ומגדיל את סיכויי היריבה לנצח.*
* *כרטיסים אדומים לקבוצה היריבה, עבור קבוצת הבית במ"ל: קבוצה שנוטה ל"לסחוט" כרטיסים אדומים נוטה להיכנס ליתרון מספרי מה שמגדיל את סיכוייה לנצח, ומקטין את סיכויי היריבה לנצח.*
* *הפרש כרטיסים אדומים בין הקבוצה לקבוצה היריבה, עבור קבוצת הבית, במ"ל: קבוצה שמקבלת יותר כרטיסים אדומים בממוצע יש סיכוי גבוה יותר שתגיע לחיסרון מספרי, בדומה לקדומים.*
* *רצף ניצחונות נוכחי: מראה האם הקבוצה בכושר "תקופה טובה", יכול גם להשפיע על המוטיבציה של הקבוצה-מומנטום חיובי.*
* *רצף ללא הפסדים נוכחי: מראה האם הקבוצה בכושר "תקופה טובה",*
* *רצף הפסדים נוכחי: מראה האם הקבוצה לא בכושר טוב ומומנטום "תקופה רעה", יכול גם להשפיע על המוטיבציה של הקבוצה - מומנטום שלילי.*
* *רצף ללא ניצחונות נוכחי: מראה האם הקבוצה לא בכושר טוב ומומנטום "תקופה רעה", אולי משקף את ביצועי הקבוצה לאחרונה יותר מאשר ביצועיה לפני כמה שבועות או חודשים. יכול גם להשפיע על המוטיבציה של הקבוצה - מומנטום שלילי.*
* *ניצחונות, שערים, תיקו, הפסדים, נקודות, הבקעות, ספיגות, הפרש שערים, אי כיבוש ואי ספיגה - עבור 5 המשחקים האחרונים בלבד. כדי לשקף את יכולת הקבוצה בתקופה שממש לפני המשחק.*

*האלגוריתמים שנבחרו עבור שיטת ההצגה הראשונה:*

*כאמור, עבור אלגוריתמים מבוססי עצים, קבוצת הדוגמאות עוברת מישקול כדי לאזן בין הסיווגים השונים.*

*עץ בודד:*

*עבור 50% מהתכונות בחרנו עץ עם הפרמטרים:*

|  |  |
| --- | --- |
| *מספר עלים מינימלי בעלה\לפני פיצול* |  |
| *עומק מקסימלי* |  |
| *מדד חשיבות* |  |
| *רווח עבור קבוצת הוולידציה* |  |

*עבור 75% מהתכונות בחרנו עץ עם הפרמטרים:*

|  |  |
| --- | --- |
| *מספר עלים מינימלי בעלה\לפני פיצול* |  |
| *עומק מקסימלי* |  |
| *מדד חשיבות* |  |
| *רווח עבור קבוצת הוולידציה* |  |

*באופן כללי נראה שמספר בינוני של תכונות-50% או 75% נתנו תוצאות עדיפות.*

*יער אקראי:*

*עבור 100% מהתכונות בחרנו יער עם הפרמטרים:*

|  |  |
| --- | --- |
| *מספר עצים* |  |
| *מספר עלים מינימלי בעלה\לפני פיצול* |  |
| *עומק מקסימלי* |  |
| *מדד חשיבות* |  |
| *מספר תכונות מקסימלי בכל פיצול* | *שורש מסך התכונות* |
| *רווח עבור קבוצת הוולידציה* |  |

*עבור 90% מהתכונות בחרנו יער עם הפרמטרים:*

|  |  |
| --- | --- |
| *מספר עצים* |  |
| *מספר עלים מינימלי בעלה\לפני פיצול* |  |
| *עומק מקסימלי* |  |
| *מדד חשיבות* |  |
| *מספר תכונות מקסימלי בכל פיצול* | *שורש מסך התכונות* |
| *רווח עבור קבוצת הוולידציה* |  |

*עבור 75% מהתכונות בחרנו יער עם הפרמטרים*

|  |  |
| --- | --- |
| *מספר עצים* |  |
| *מספר עלים מינימלי בעלה\לפני פיצול* |  |
| *עומק מקסימלי* |  |
| *מדד חשיבות* |  |
| *מספר תכונות מקסימלי בכל פיצול* | *שורש מסך התכונות* |
| *רווח עבור קבוצת הוולידציה* |  |

*ביער נראה שיותר תכונות נותן תוצאות עדיפות*

*יער אקראי מאוד (extra):*

*עבור 100% מהתכונות בחרנו יער עם הפרמטרים:*

|  |  |
| --- | --- |
| *מספר עצים* |  |
| *מספר עלים מינימלי בעלה\לפני פיצול* |  |
| *עומק מקסימלי* |  |
| *מדד חשיבות* |  |
| *מספר תכונות מקסימלי בכל פיצול* | *מסך התכונות* |
| *רווח עבור קבוצת הוולידציה* |  |

*האלגוריתם אכזב ביחס ליער אקראי רגיל, לא ראינו טעם בבחירת יערות נוספים לקבוצת המבחן חוץ מהאחד הזה.*

*גרדיאנט בוסט:*

*עבור 100% מהתכונות, מסווג ראשוני ‘zero’ (מסווג האפס) בחרנו בפרמטרים:*

|  |  |
| --- | --- |
| *מספר עצים* |  |
| *מספר עלים מינימלי בעלה\לפני פיצול* |  |
| *עומק מקסימלי* |  |
| *משקל חשיבות לכל עץ* |  |
| *אחוז הדוגמאות ללימוד כל עץ* |  |
| *מספר תכונות מקסימלי בכל פיצול* | *מסך התכונות* |
| *רווח עבור קבוצת הוולידציה* |  |

*עבור 100% מהתכונות ועץ ראשוני שלנו כלומר העץ המוצלח ביותר עבור 100% מהתכונות, בחרנו בפרמטרים:*

|  |  |
| --- | --- |
| *מספר עצים* |  |
| *מספר עלים מינימלי בעלה\לפני פיצול* |  |
| *עומק מקסימלי* |  |
| *משקל חשיבות לכל עץ* |  |
| *אחוז הדוגמאות ללימוד כל עץ* |  |
| *מספר תכונות מקסימלי בכל פיצול* | *מסך התכונות* |
| *רווח עבור קבוצת הוולידציה* |  |

*עבור 90% מהתכונות ועץ ראשוני שלנו כלומר העץ המוצלח ביותר עבור 90% מהתכונות, בחרנו בפרמטרים:*

|  |  |
| --- | --- |
| *מספר עצים* |  |
| *מספר עלים מינימלי בעלה\לפני פיצול* |  |
| *עומק מקסימלי* | *לא מוגבל* |
| *משקל חשיבות לכל עץ* |  |
| *אחוז הדוגמאות ללימוד כל עץ* |  |
| *מספר תכונות מקסימלי בכל פיצול* | *מסך התכונות* |
| *רווח עבור קבוצת הוולידציה* |  |

*עבור 75% מהתכונות ועץ ראשוני שלנו כלומר העץ המוצלח ביותר עבור 75% מהתכונות, בחרנו בפרמטרים:*

|  |  |
| --- | --- |
| *מספר עצים* |  |
| *מספר עלים מינימלי בעלה\לפני פיצול* |  |
| *עומק מקסימלי* |  |
| *משקל חשיבות לכל עץ* |  |
| *אחוז הדוגמאות ללימוד כל עץ* |  |
| *מספר תכונות מקסימלי בכל פיצול* | *מסך התכונות* |
| *רווח עבור קבוצת הוולידציה* |  |

*אלגוריתם ה*gradient boost *הראה ביצועים טובים ויחסית יותר יציבים ולכן יש לנו ציפיות גבוהות ממנו.*

*רגרסיה לינארית:*

*עבור רגרסיה לינארית מסוג ridge, בחרנו עבור 100% מהתכונות את הפרמטרים:*

|  |  |
| --- | --- |
| *שיטת פתרון* |  |
| *\*חוזק רגולרזציה* |  |
| *\*\*יירוט קבוצת הדוגמאות – מה זה בדיוק?* |  |
| *נרמול – האם רלוונטי? כי כבר עשינו להכול* |  |
| *רווח עבור קבוצת הוולידציה* |  |

*\*שולט על ה-tradeoff בין שונות ו-bias.*

*עבור רגרסיה לינארית מסוג elastic net בחרנו עבור 50% מהתכונות את הפרמטרים:*

|  |  |
| --- | --- |
| *חוזק רגולרזציה* |  |
| *\*ענישה* |  |
| *\*\*יירוט קבוצת הדוגמאות – מה זה בדיוק?* |  |
| *נרמול – האם רלוונטי? כי כבר עשינו להכול* |  |
| *רווח עבור קבוצת הוולידציה* |  |

*\* מה זה ענישה -*

*באופן כללי התוצאות היו פחות טובות מאשר עצים אבל יציבות יחסית. אמנם בדרך כלל מפסיד אבל פחות מאשר מהמר רנדומלי.*

*רגרסיה לוגיסטית מולטי:*

*עבור* solver *מסוג saga, עם כל התכונות בחרנו את הפרמטרים:*

*(1.0, False, 0.4, 'weighted')*

*שנתנו רווח של 0.028667776852622815*

*עבור* solver *מסוג saga, עם כל התכונות בחרנו את הפרמטרים*

*(1.0, True, 0.6, 'weighted')*

*שנתנו רווח של 0.010184623863323222*

*עבור* solver *מסוג saga, עם 90% מהתכונות בחרנו את הפרמטרים*

*(0.4, True, 0.8, 'weighted')*

*שנתנו רווח של 0.009203220433092735*

*רוב ה*solvers *והפרמטרים לא הגיעו לתוצאות טובות כל כך בחרנו רק את היותר טובים.*

*בגדול נראה שסינון תכונות פוגע ברווחים.*

*רגרסיה לוגיסטית סינגל:*

*עבור פוטר מסוג lbfgs עם 25% מהתכונות בחרנו את הפרמטרים:*

*(0.1, True)*

*שנתנו "רווח" (הפסד) של -0.00707562568008703*

*יש לציין שבניגוד לאלגוריתמים אחרים, אלגוריתם זה הראה יציבות רבה כמעט בכל שילובי הפרמטרים השונים, ושדווקא סינון תכונות שיפר הביצועים (אם כי עדיין אין רווח).*