# בדיקת הקשר בין אירועים במשחק כדורגל לניצחון והפסד קבוצת הבית

פרוייקט בבינה מלאכותית

236502

מגישים:

**מבוא:**

חיזוי תוצאות משחק כדורגל הינו משאלה של כל אוהד כדורגל או מהמר, אך גם משאת ליבם של לא מעט מתכנתים ודי בחיפוש שטחי באינטרנט בכדי למצוא מאגרי נתונים הפרושים לציבור הרחב על מנת לנסות ולהפיק תחזית עבור תוצאת המשחק. יחד עם זאת ככל הנראה שאין אלגוריתם מושלם היודע לספק תחזית בהסתמך אך ורק על סטטיסטיקות קודמות. אילו היה כזה הרי שהמחזיק במידע היה יכול להתעשר בין לילה באמצעות חיזוי טופס טוטו או ווינר עם יחס גדול.

במאמר זה ננסה למצוא קשר בין האירועים במשחק עצמו ובין התוצאה של קבוצת הבית. ניבוי זה נראה לנו יומרני פחות מאחר והוא אינו מתבסס על נתונים וסטטיסטיקות קודמות אלא בעיקר על מידע שברובו נוצר תוך כדי מהלך המשחק. יחד עם זאת הוא מאוד מעניין. ננסה לגלות מהן התכונות הדומיננטיות והאם תכונות שנראות שוליות הינן בעלות משקל משמעותי על תוצאת המשחק. על מנת לגלות ולבחון את הקשר אספנו את הdata הגולמי הפזור ברחבי האינטרנט ועיבדנו אותו לכדי data שמכיל מספר רב של תכונות בכל משחק. את התכונות הנ"ל נחקור לעומק באמצעות אלגוריתמי למידה מסוגים שונים. ננסה להבין מהי הדרך הנכונה להציג את התכונות שחילצנו ואילו מהן אכן קשורות בקשר חזק לתוצאת המשחק.

לאחר חיפוש מצאנו המון מידע באתר Kaggle אשר מכיל מאגר מידע רב. המידע כפי שציינו לא היה מלוטש והיה צורך בעיבודו. במאמר ננסה להבין בין היתר מדוע בחרנו בצורות הצגה מסוימות עבור התכונות שאספנו. חלק זה הינו בעל משמעות רבה שכן חילוץ מידע והצגתו בצורה נכונה יכול להשפיע על רמת הדיוק של האלגוריתם.

**זיהוי תכונות**

התכונות שחילצנו הינן התכונות הבאות:

התאריך שבו המשחק התרחש, סוג הליגה, מספר עונה, המדינה בה התרחש המשחק, קבוצת הבית, קבוצת החוץ, מספר שערים שהבקיעה קבוצת הבית, מספר השערים שהבקיעה קבוצת החוץ, תוצאות ההימורים עבור ניצחון לקבוצת הבית, תוצאות ההימורים עבור תיקו, תוצאות הימורים עבור ניצחון קבוצת החוץ, הזמן בו האירוע התרחש.

לאלו נוספו:

11- Attempt(shot)  
2- קרן  
3- עבירה  
4- כרטיס צהוב  
5- כרטיס צהוב שני  
6-כרטיס אדום ישיר   
7- חילוף  
8- Free kick won  
9- נבדל  
10- נגיעת יד  
11- פנדל   
12 – כדור עובר לקבוצה השנייה  
13 - כדור כושל  
14-כדור חוץ   
15-גול עצמי

כל אירוע שכזה הינו כפול למעשה מאחר והוא יכול לקרות לטובת קבוצת הבית ולטובת קבוצת החוץ. ניתן היה לכווץ את התכונות ולא להכפילם בעמודות נוספות על ידי חיסור בין מספר הפעמים בהם האירוע קרה עבור קבוצת הבית ומספר הפעמים בהם הוא קרה עבור קבוצת החוץ אך העדפנו להכפיל את התכונות ולהכניס כתכונה את מספר הפעמים בהם אירע אירוע i עבור קבוצה j. כלומר חילצנו את מספר הפעמים בהם אירע האירוע לטובת קבוצת הבית והצגנו זאת כתכונה ואת מספר הפעמים בהם האירוע היטיב עבור קבוצת החוץ. בצורה זו נוכל לבדוק האם תכונות דומיננטיות עבור צד אחד הינן דומיננטיות גם עבור הצד השני הרחבה על כך תינתן בהמשך.

מהתכונות הקודמות החלטנו להסיר את התאריך מאחר והוא לא נראה לנו כגורם שניתן לזהות באמצעות אלגוריתם למידה כדומיננטי מאחר ומדובר בערכים השונים אחד מהשני בצורה מגוונת מידי ביחס לנתונים. מסיבה דומה לא שמרנו את שמות הקבוצות אלא הסתפקנו בייצוג דליל של קבוצת בית וקבוצת חוץ. כמו כן הסרנו את המדינה מאחר והיא כלולה בסוג הליגה.

מרבית התכונות שהשארנו הינן מספריות בדידות מאחר ומדובר **במספר הפעמים** שהאירוע התרחש. מאידך, קיימות גם מעט תכונות קטגוריאליות דוגמת "סוג הליגה" ואף תכונות מספריות אשר הינן למעשה קטגוריאליות דוגמת "מספר העונה" שהרי אין באמת ערך למספר כמספר והוא אינו חשוב מפני הסדר שהוא מייצג. בנוסף לאלו קיימות תכונות מספריות רציפות דוגמת "תוצאת ההימורים עבור קבוצת בית\חוץ\תיקו".

**פיצול ה-DATA**

החלטנו לחלק את הדאטא כך:

Train 60%

Validation 20%

Test 20%

בחרנו גודל יחסית גדול לקבוצת הוולידציה כי רצינו שהבדיקות עליה ידמו בצורה קרובה ככל הניתן את הבדיקות על קבוצת המבחן.

**בדיקת ביצועים**

כדי להעריך את טיב קבלת ההחלטות שביצענו, כמו בחירת המתודות בהן אנו עושים outlier detection, scaling ו-feature selection וקביעת פרמטרים שונים לאלגוריתמים בהם השתמשנו, החלטנו להשתמש במסווג ולבדוק את ביצועיו (אחוזי ה-accuracy) על ה-data שעבר עיבוד על ידינו, אשר מאומן על ה-train data ונבדק על ה-validation data.

החלטנו לבחור במסווג שביצועיו על ה-data הראשוני שלא עבר עיבוד לא גבוהים מידי, כדי שנוכל לראות שינויים יחסית משמעותיים כאשר נריץ אותו על ה-data המעובד. כמובן שלצורך הבחירה הסרנו את התכונות של תוצאת המשחק עבור קבוצת הבית וקבוצת החוץ להם קיים קשר ישיר עם החיזוי (לשם המחשה כאשר בדקנו את ה- accuracyעם התכונות הנ"ל התקבלה התוצאה 1 עם random forest ו- 0.999 עם svm).

אולם ראשית עלינו להתמודד עם השאלה מהי הצלחה. מאחר ולמשחק קיימות 3 אפשרויות שונות ואנו חוקרים רק אחת מהן לא יהיה נכון לומר שמעל 50 אחוז נחשב להצלחה בדיוק כשם שאלגוריתם החוזה את מספר הפעמים בהם לא נקבל 6 בקובייה ומסמן שלא קיבלנו כל הזמן לא יחשב להצלחה מאחר וההסתברות לטעות היא 1/6. על מנת להעריך את טיב הביצועים החלטנו להריץ את הdata שלנו עם עמודות רנדומליות ולבחון את טיב הביצועים שלו.

לצורך בחירת מסווג אשר יעזור לנו רק להעריך את ה-data preparation החלטנו לבדוק את random forest ו-svm בהרצה עם עמודת ערכים רנדומלית התקבלו התוצאות הבאות:

|  |  |
| --- | --- |
| **מסווג** | **accuracy** |
| Random forest | 0.557 |
| svm | 0.557 |

כלומר עבור נתונים רנדומליים אחוז הדיוק הוא קצת מעל 50 אחוז. אם נצליח להשיג אחוזים משמעותיים יותר ככל הנראה שהמסווג עובד. כמות הנתונים הכוללת שאספנו הינה כ-10,000 משחקים ולכן חוק המספרים הגדולים תופס למקרה שלנו. לצורך המחשה עבור הרצה רנדומלית לחלוטין עם מסווג טרינארי ועמודות רנדומליות התקבל דיוק של 0.442.

כעת נעבור לבחירת המסווג הראשוני אשר איתו נבצע בדיקה ראשונית עבור עיבוד הדאטא. בהשוואה בין המסווגים הנ"ל התקבלו התוצאות הבאות:

|  |  |
| --- | --- |
| **מסווג** | **accuracy** |
| Random forest | 0.695 |
| svm | 0.557 |

החלטנו לבחור במסווג svm שביצועיו על ה-data הראשוני היו דומים למסווג רנדומלי ולבדוק אם אכן ביצועיו משתפרים לאחר עיבוד ה-data.

**עיצוב תכונות**

כפי שציינו לעיל ניתן לייצג את הdata שלנו בצורה כפולה כאשר פעם האירועים יספרו עבור קבוצת הבית ופעם עבור קבוצת החוץ (תצוגה א). אך ניתן לייצגם בצורה של חיסור התכונות בהם קרה הדבר לטובת קבוצת החוץ מהפעמים בהם האירוע התרחש לטובת קבוצת הבית (תצוגה ב). נשווה בין התוצאות על מנת לזהות האם מתקבלת השפעה משמעותית על הדיוק.

לאחר השוואה התקבלו תוצאות זהות ברמת הדיוק בין שני המקרים ולכן בחרנו להמשיך עם תצוגה א בכדי לראות האם תכונות דומיננטיות עבור קבוצה אחת הינן דומיננטיות גם עבור קבוצה אחרת אך נמשיך להשוות ביניהן בהמשך.

שמירת תוצאות ראשוניות: