אמיר וולפנזון 300339785, אור שריר 201314796 , אדם ברטאש 206321440, ישי שפירא203016217

פרויקט סקרנות - חלק א' - האם ננצח משחק כדורגל וכמה נרוויח מכך?

1. א. קלט: נתונים לנו קבוצות מ-6 ליגות באירופה (ליגה ספרדית, ליגה אנגלית, ליגה איטלקית, ליגה צרפתית, ליגה גרמנית וליגה הרוסית), והתוצאות של הקבוצות מהליגות האלה בין השנים 2014 ל-2019. בקובץ אחד נתון סיכום העונה של כל קבוצה בכל אחת מהשנים, ובקובץ האחר יש תיאור של כל קבוצה פר משחק בכל אחת מהעונות.
הפרמטרים הם: הליגה, מספר גולים שהיו אמורים להיכנס עפ"י מדד expected goals, מספר גולים שהקבוצה היתה אמורה לספוג (.(expected goals against
פלט: נרצה לדעת בהינתן משחק של קבוצה, האם תנצח, תפסיד או תסיים בתיקו ומה תרוויח מכל מצב שכזה.

ב. תחומים:

- גולים צפויים למשחק (מ-0 ועד 8 למשחק).
- גולים צפויים שהקבוצה סופגת (מ-0 ועד 8 למשחק).
- מספר הליגה (מספור מ-1 עד 6 עפ"י ליגה ספרדית, ליגה אנגלית, ליגה איטלקית, ליגה צרפתית, ליגה גרמנית וליגה הרוסית).

2 הראשונים הם ערכים נומריים, לפיכך, הפכנו אותם לבינים כאשר יש קשר בין 2 ערכים עוקבים. למשל, יש קשר בין xg=3, לבין .xg=4

ערך המטרה: 2 אם הקבוצה הביתית תנצח, 1 אם קבוצה שבחרנו תסיים בתיקו, 0 אם קבוצה שבחרנו תפסיד כאשר מדובר על ערכים בעלי משמעות ולא רק מתארי מצב (הערך שנקבל מנצחון הוא 2, מתיקו הוא 1 ומהפסד הוא 0).

תוצאות רצויות: אנחנו נרצה לחזות האם משחק המתקיים בזמן אמת אכן הסתיים בהתאם למה שחזינו או לא.

ג. השתמשנו במאגר המידע של Football Data: Expected Goals and Other Metrics מתוך Kaggle.

2. א. במקרה שלנו אנחנו רוצים למצוא פרמטר יחיד בהינתן משחק: האם הקבוצה שבחרנו תנצח, תפסיד או תסיים בתיקו ומה הערך שתקבל מכך.

ב. את הדאטה שלנו נחלק לנתונים שמעניינים אותנו. 80% מהנתונים ישמשו אותנו לקובץ training set, ו-20% מתוכו ישמש אותנו ל-test. השלב הסופי יהיה לבדוק האם משחק בזמן אמת בטלויזיה (של קבוצות המופיעות אצלנו בדאטה סט) הסתיים בתוצאה שחזינו או לא. חישוב ה-PRIOR:

24580 משחקים בין השנים 2014-2019, מתוכם היו 9189 נצחונות של קבוצה (באופן דומה, גם 9189 הפסדים של קבוצה). לפיכך:

P(H_wins) = P(H_lose) = 0.37.

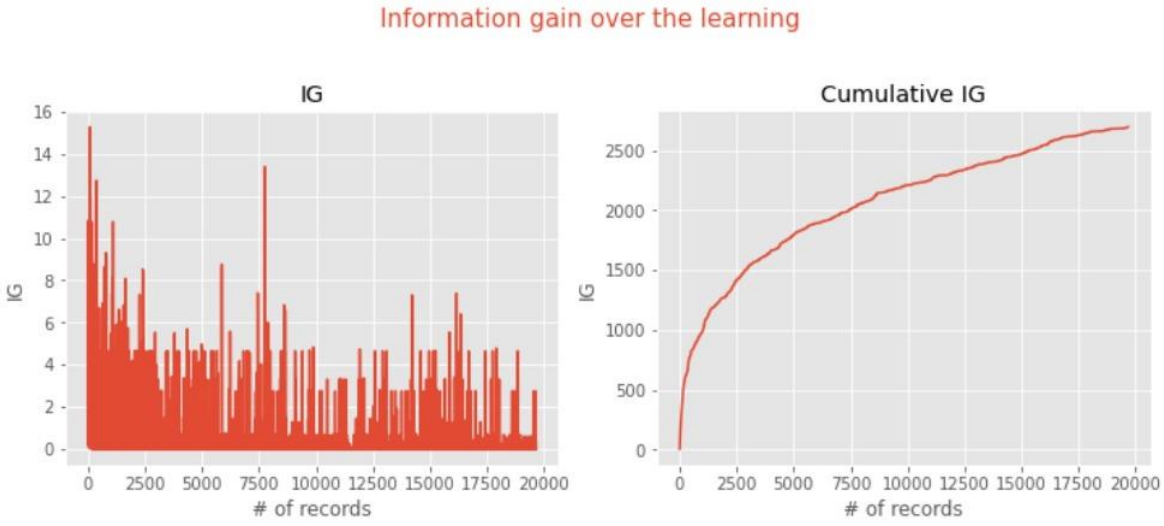
24580 משחקים בין השנים 2014-2019, מתוכם היו 6202 תוצאות תיקו:

P(H_draw) = 0.26.

3. א. בכל איטרציה עברנו על שורה חדשה בtrain data ועדכנו את מטריצת ההסתברויות בערך של נקודת הדגימה, בנוסף עדכנו את הסביבה הקרובה של ערכי ה-xG ו-xGA, לא עדכנו ערכים דומים לליגה שקיבלנו בנקודת דגימה מכיון שזהו ערך קטגוריאלי שאין בו דמיון בין הערכים השונים - הליגות השונות.

מצורף גרף הממחיש את השיפור ב-information gain. ניתן לראות שככל שעוברים על יותר שורות, לומדים פחות. הגרף מימין ממחיש את ערך הלמידה הקומולטיבי.

```
Out[780]: Text(0, 0.5, 'IG')
```



ב. הנתון שממנו אנחנו לומדים הכי הרבה הוא תוצאת תיקו, כפי שניתן לראות בגרף ה-information gain למטה, במידה והיינו צריכים לבחור רק 2% מהדאטה שלנו וללמוד רק ממנו היינו לוקחים שורות בהם תוצאת המשחק היה תיקו

```
Out[781]: <matplotlib.axes._subplots.AxesSubplot at 0x24ca6cadb50>
```