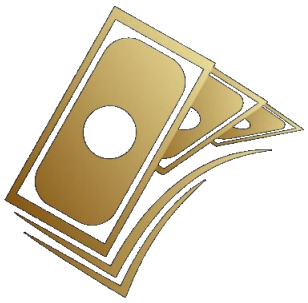


BETTER BETTOR

SPORTS BETTING MADE EASY

Table of Contents

Abstract

Sports betting is a popular and lucrative activity that attracts billions of fans worldwide, but it also poses significant challenges for rational decision-making

and addiction prevention. This project examines how machine learning, a branch of artificial intelligence that can analyze large amounts of data and identify patterns, can help sports bettors make better choices and improve their outcomes. It also explains how the sports betting industry works, how the bookmakers gain an advantage over the bettors, and what types of bets can be placed. The project argues that machine learning can help sports bettors overcome the limitations of human intuition and emotion, as well as provide them with more accurate and reliable predictions based on data and statistics. The project also discusses the potential benefits and drawbacks of using machine learning for sports betting, and suggests some directions for future research.

Introduction

We all know that professional sports capture the hearts of billions of fans worldwide, yet even though watching the sport is entertaining enough, sports betting has become quite a popular activity among fans. But as with casino gambling or any other form of gambling, the odds are always in favor of the house... by a significant margin. In fact, according to recent reports, the sports betting industry in the United States has grown with relative speed, with many states legalizing it in the pastime and even more in the process of doing so. In 2022, sports betting revenue was forecasted to reach 7.56 billion U.S. dollars [**1**]. But despite this growth, sports bettors often lose due to irrational decision-making and addiction. According to a recent article by Kindbridge Behavioral Health, sports betting is often motivated by factors such as the desire to make the sport more interesting, the excitement and thrill of betting, and the need to deal with risk [**2**]. Many sports bettors do not bet using any information or data, but rather rely on their emotions and gut feelings. This can lead to irrational decision-making and even addiction. So what can be done about this? Of course, the obvious answer is: don't gamble, but that isn't so easy to implement. What if there was a way to bet more rationally and not rely so much on guts or instinct? Enter: machine learning. In recent years, machine learning has emerged as a powerful tool that can be used to make better decisions in sports betting. By analyzing large amounts of data and identifying patterns, machine learning algorithms can improve the accuracy of sports betting gambles and the health our bank accounts as a result.

How exactly does the house win in sports betting?

It is no surprise that there are many ways for the house to win in sports betting. It all starts with the market being initiated by a bookmaker, who accepts and pays off bets at agreed-upon odds. However, just like in the game of Roulette, even the house is not certain of the outcome, so their advantage must come from elsewhere. In Roulette, the house advantage lies in the risk to payout ratios relative to the statistical probability of each event occurring (i.e. expected value). While bookmakers can also implement this property to some extent by

setting appropriate odds, they are not certain of the probabilities either, as sporting events have an unpredictable nature. Instead, bookmakers apply what is called vigorish, which is a commission charged for laying bets. This means the odds no longer correspond with the true probabilities of the events occurring, creating a profit margin for the bookmaker. Additionally, bookmakers typically balance their book by adjusting the odds even further and indirectly impacting the amount of money bet on each event (people will be discouraged to bet if the odds aren't "good enough"), in order to guarantee a profit regardless of the outcome. On a larger scale, bookmakers may also reduce their risk by spreading it to other bookmakers, similar to insurance and reinsurance companies, but typically this is only done if they are facing fewer betters betting large amounts.

How can players win in sports betting?

Winning from the players perspective in sports betting is simple. You place your bets and if the outcome of the event is as you predicted, you receive a payout relative to the odds agreed upon between you and the bookmaker. Negative odds imply that the event being bet on is favored and thus such expected results payout less than unexpected ones, meaning a player would have to wager more money on an expected outcome to receive a similar payout as an unexpected one (like an upset win for which the underdogs are given positive odds). This may seem obvious or very logical, however the complexity arises from the variation of bets that can be placed. The simplest bets are money-line bets, where you bet on a team to outright win, lose, or draw. Spread betting involves the spread, which not only identifies the favored team, but also how much they are expected to win by. With this type of betting, players can bet on the favored team to win by more than the spread, or on the underdog to either lose by less than the spread or win outright. Total (over/under) bets are similar to spread bets, but reflect the total score of the game combining both the home and away team scores. Parlay bets require the player to combine bets from multiple games into one big bet. Here, the payouts are significantly higher, but correspond to the higher unlikelihood of all events occurring as predicted. If at least one of the bets included in the parlay doesn't win, the entire bet is lost. There are other more complex bet types, but these are the most common and important ones to understand.

So what's the goal here?

With this in mind, our goal here is to create predictive models for various sports and leagues in order to assess the bookmaker's odds and find the most profitable betting opportunities. It's important to note that there is never any certainty in this endeavor, as these are only predictive models and will always have some error involved. Therefore, ideally, we want to minimize this error and maximize profits.

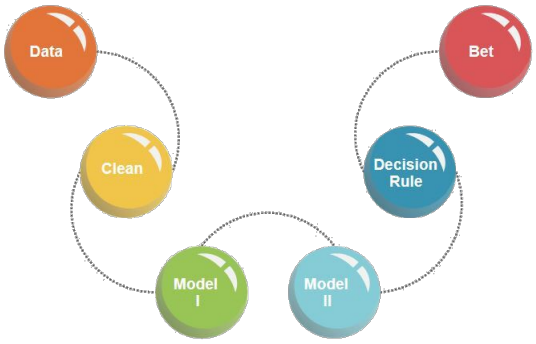
Data Description

We are working on a project that involves analyzing data from Hockey Reference and MoneyPuck, two websites that provide statistics and insights on hockey games and players. Our goal is to create objects for the teams that we can use for further analysis and modeling. To do this, we need to clean the data by selecting the columns that have the most value for our project (i.e. home team, away team, home goals, away goals, date played) and filtering the data by the situations that are relevant for our research question, including web scraping sports book data from Covers.com This includes whatever odds the bookies set on the games we are modeling across various props such as money-line, over/under, etc. We are not using all the data available from these sources, because some of it may introduce noise and bias into our model. By cleaning the data, we hope to improve the accuracy and reliability of our results.

Methodology

Who ya gonna call?

Creating a predictive model for sports betting can be quite challenging at first due to the variables involved. To make the process more manageable, we have divided the model into two parts: predictions and odd evaluations. First, we determine our predictions (what we claim will happen), then evaluate the book-maker's odds by calculating the corresponding probabilities (what they claim will happen) and use the information to make a decision on which events to bet on and which to avoid. A visualization of the two-step model can be seen below:



This diagram represents the entire pipeline of what is to be done with the incoming data set in order to result in making the bets we would like to make. We start by defining what we want our final data set to look like and then proceed to clean it. Once cleaned, we move along with our modeling. The first model handles the issue of predicting who will win, lose, or draw, while the

second model predicts the margin of said victory or defeat. Based off of these model results, we can assess the bookmakers odds and find inconsistencies with their evaluation of the game and potentially take advantage. But in order to understand exactly how we will achieve our goal here, a closer look into the data modeling is now needed.

Thank you Arpad Elo!

In making a money-line bet, our aim is to predict the outcome of an event, not the score. Therefore, we can use the Elo Rating System to power rank the teams. The Elo Rating System determines the relative quality of a team, and the difference in ratings between two teams serves as a predictor of the outcome of a match between them (i.e. teams that are higher rated than others are more likely to win). We will explain how the system works below, but before we do that, a few important properties regarding the system must be mentioned:

- The ELO Rating System is self-correcting.
  - Teams whose ratings are too low or too high will gain or lose rating points until the ratings reflect their true playing strength.
- ELO ratings are comparative and only valid within the rating pool in which they were calculated.
  - We should only compare ELO ratings within the same league or competition.

However, if we want good results to come from this model, we must wait to observe enough games to ensure that the 'self correction' occurs and the teams reflect their true ratings. But if we want to make predictions from the start of a given season, we cannot afford to wait (or incorrectly predict games) in the short run. As a result, a key observation is made. Because of the 'self-correcting' property of the ELO Rating System, we say that the ELO Rating System is a Markov chain following the Markov property:

$$P(X_n = j|X_{n-1} = i) = P(X_n = j|X_{n-1} = i, X_{n-2} = i_{n-2}, ..., X_0 = i_0) \quad (1)$$

This means that the probability of a team's rating being j now (at time n) given that it was i at the end of last season (at time $n-1$) is the same as the probability of the team's rating being j given all previous seasons up to the first one (at time 0). Why is this useful? Well, because of this notion of the ELO Rating System being memory-less, instead of arbitrarily assigning initial ELO ratings to each team in the beginning of the season we want to forecast, we can instead get accurate predictions by calculating the ELO ratings of the season prior (since by the Markov property these calculations are the same). In fact, we can theoretically use data from as many seasons back in time as we wish, however in practice, this would not yield the best results as ratings would face

rating inflation. This can occur in any sport where if we go back in time enough, the relative skill of the players changes due to various factors causing the ELO system to inaccurately represent the true skill of teams and players.

Applying the Elo system.

By applying the Markov property, we now know that instead of having to start the 'self-correction' process from the first ever season played (n seasons ago), we can start at the season before the present one (season $n-1$). Therefore, we assign every team in the league/competition an initial rating of 1500 since all teams are theoretically at equal strength at the beginning of the (first ever) season. Given each team's ELO rating, we can update the ratings of each team as they face new opponents to get an idea of their relative strength. To do this, we first calculate the probability, E_H, of the home team winning against the visiting team given their respective ELO ratings as follows:

$$E_H = \frac{1}{1 + 10^{-(R_H - R_V + Home)/400}} \tag{2}$$

where R_H is the Elo rating of the home team, R_V is the Elo rating of the visiting team, and $Home$ is the home advantage coefficient (i.e. the proportion of games won at home across the entire league). Since something is certainly expected to happen (i.e. $E = 1 = E_H + E_V$), we also get that $E_V = 1 - E_H$ as a result. Once we have calculated the expected scores and observed the actual result of the event, we can proceed to calculating the change in Elo rating for both teams. Suppose the home team with rating R_H is expected to score E_H points but actually scores S_H points. Then the formula for updating that team's rating is:

$$R'_H = R_H + K \cdot (S_H - E_H) \tag{3}$$

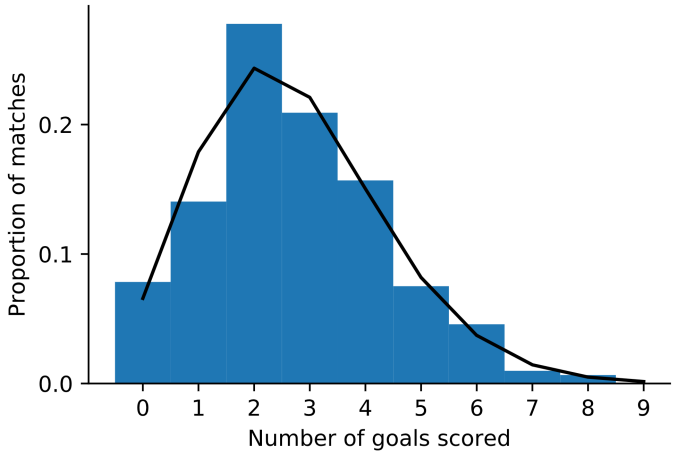
where R'_H is the updated rating, R_H is the previous rating, and K is the adjustment factor (usually set to 32). The same logic can be applied to determine the new ELO rating of the losing team. Finally, we can repeat this process for all games that occur during the season and use the resulting ratings to make predictions in the next season.

It's all about the margin.

Now that we have the ELO model to evaluate relative skill, we can use the ELO ratings of two teams and calculate the difference between them and use that as a predictor for the difference in goals scored. For example, if Team A has an ELO rating of 1500 and Team B has an ELO rating of 1400, the difference is 100. This means that Team A has a 64% chance of winning and is expected to score 1.07 more goals than Team B. By using the difference in ELO ratings as

a predictor, we can create a regression model that can estimate this difference in goals scored based on the difference in ELO ratings. This can help us to evaluate the performance of teams and predict the outcomes of future matches.

Now that we can predict future expected goals, how do we use this to get an idea of actual future goals? To answer this, a key observation is made by considering the visual below:



We can see that goals scored over time follow a Poisson distribution. This means that once we get the expected goals for a team (i.e. the average number of goals they will in the game), we draw probability calculations for actual goals as follows:

$$P(Goals = G) = \frac{e^{-(xG)} \cdot (xG)^G}{G!}$$

where G is the actual number of goals and xG is the expected goals that the team will score. We can use these probabilities and compare them to the ones the bookmakers are offering.

Now we have a complete system of models for predicting game outcomes. The next steps will focus on the second half of the pipeline, where we draw probabilities and evaluate odds to determine which games to bet on and which ones to avoid.

To catch a bookmaker.

Now that we have finalized our predictions we will compare them to the odds offered by bookmakers to find potential opportunities for profit. The bookmakers

do not release their own probabilities, but we can infer them from the odds they offer. These inferred probabilities are called implied probabilities. To calculate the bookmaker's implied probability, we need to know the odds they are offering for a particular event in the event space. Depending on where the bookmaker is located, odds may be reported in different ways. To avoid confusion, we will show the calculation for all three forms of odds, but our convention will be based on American odds. To convert fractional odds to decimal odds, the following formula is used:

$$\text{Decimal Odds} = \text{Numerator}/\text{Denominator} + 1 \tag{14}$$

If we are directly given the decimal odds, we can then convert those to American odds as follows:

$$\text{American Odds} = \begin{cases} -100/(decimal - 1), & 1.00 \leq decimal \leq 2.00 \\ (decimal - 1) \cdot 100, & decimal > 2.00 \end{cases} \tag{15}$$

Now that we have a clear understanding of what odds the bookmaker has offered, we will convert these odds to implied probabilities as follows:

$$P_{Implied}(odds) = \begin{cases} \frac{|odds|}{(|odds|+100)}, & odds < 0 \\ \frac{100}{(odds+100)}, & odds > 0 \end{cases} \tag{16}$$

To calculate the bookmaker's implied probability, we take the positive value of the odds regardless of the original sign. Then, based on the sign of the odds, we use the appropriate formula to convert the positive odds to an implied probability (see example below).

$$\text{Celtics at} -310 \implies P_{Implied}(-310) = \frac{|-310|}{(|-310|+100)} = 0.756$$

$$\text{Heat at} +245 \implies P_{Implied}(245) = \frac{100}{(245+100)} = 0.289$$

It is important to notice that for a given event, the sum of the implied probabilities from the bookmakers is greater than 1. This is because bookmakers include a margin for themselves to ensure a profit (recall vigorish). Therefore, using the implied probabilities, we must establish a decision rule to evaluate which odds to bet on and which ones to avoid. This will allow us to effectively take the most valuable risks and not waste money. Since, the implied probability is directly proportional to the ratio of the amount of money we wager to the amount of money we can potentially win:

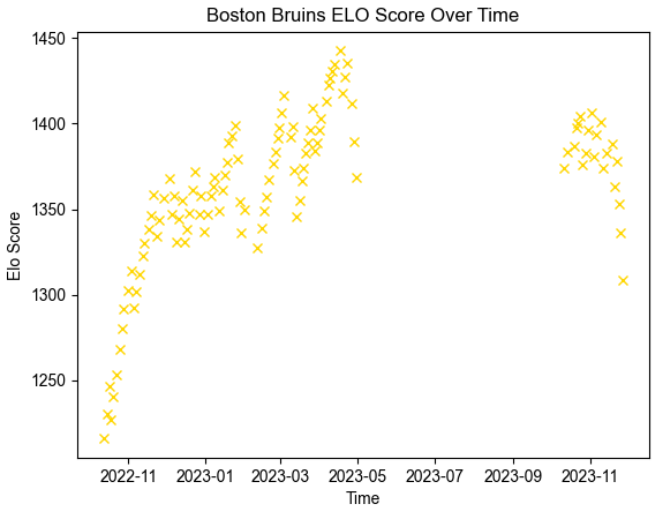
$$\text{Implied Probability} = \frac{Wager}{\text{Wager + Payout}} \tag{17}$$

Therefore, we are looking for scenarios where we can maximize the potential payout. This is because we can adjust our bet size or overall betting strategy, but we cannot change the odds set by the bookmakers. Therefore, our profits should come from optimizing potential payouts and not on having to alter our wager size. To achieve this, we observe that as we increase the potential payout, the denominator also increases and thus the implied probability decreases. What this tells us is that when we maximize the potential payout whilst holding our wager constant, we end up minimizing the implied probability. But how can we tell which implied probabilities are "low enough" and which ones aren't? We can use our predicted probabilities as a comparison since they reflect what true or fair odds should be. As a result, we get our decision rule for maximizing profits:

$$\text{Decision Rule(Bet)} = \begin{cases} 1, & \text{if Predicted Probability} > \text{Implied Probability} \\ 0, & otherwise \end{cases}$$

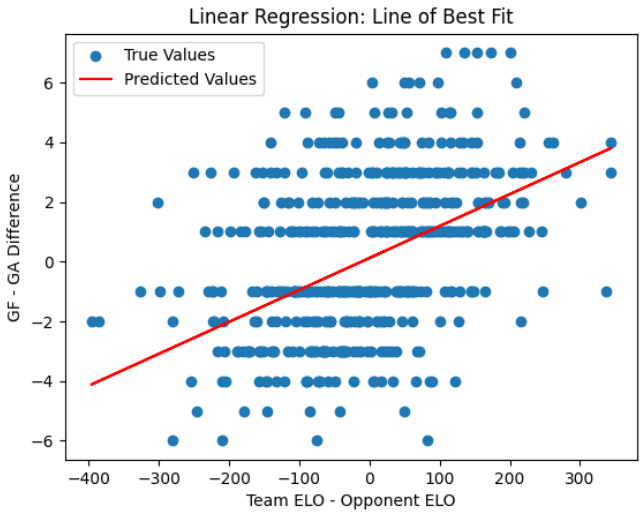
Results

We can see how the ELO model learns about a team's true rating over time by looking at the how the Bruins have played since the 2022-23 season until now:



In the case of the Bruins, we can see how they had an amazing season last year, coming just short in the playoffs after losing to Florida in game 7.

With this, the linear regression model can be visualized as follows:



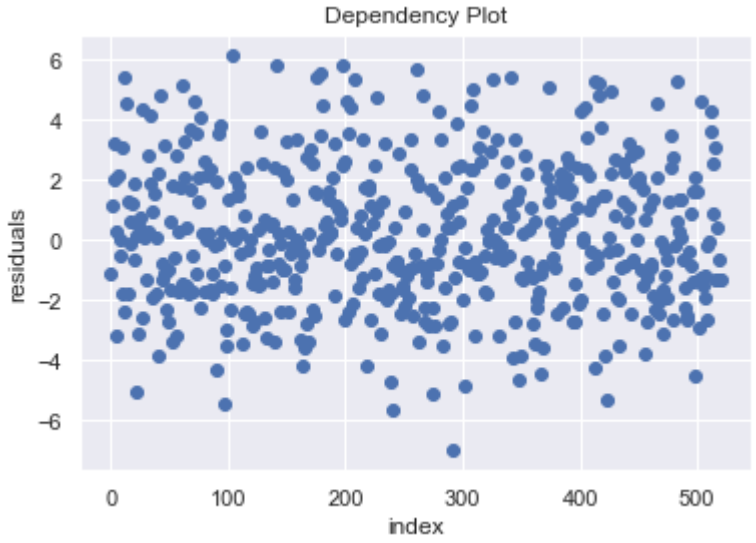
where the equation of the line is $y = 0.1284 + 0.0095x$ with an R^2 value of 0.169. This shows that for every 105 difference in ELO rating between two teams, the higher rated team is expected to score 1 more goal.

A constant variance plot of the results is then derived as follows:

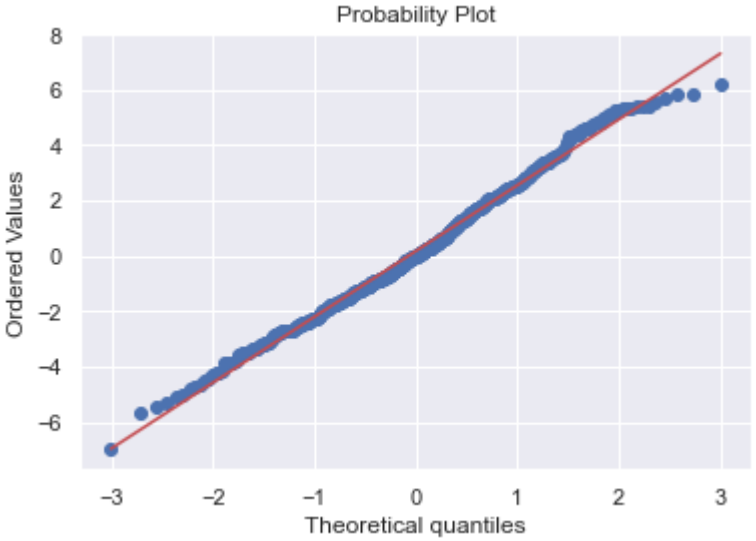
This shows that the variance in the model is not constant and thus there are

other variables for which the liner model is not sufficient in predicting the spread.

The dependency plot showed that the residuals were in fact random which means that the model assumptions do in fact check out.



Finally, the probability plot shows that the errors are, for the most part, evenly distributed.



Discussion

The linear regression model shows that the ELO rating difference between two teams has a positive and significant effect on the expected goal difference in favor of the higher rated team. However, the R^2 value of 0.169 indicates that the model only explains about 17% of the variation in the goal difference. This suggests that there are other factors that influence the outcome of a hockey game, such as injuries, penalties, home advantage, and luck. A more complex model that incorporates other variables might be able to capture the dynamics of the game better and provide more accurate predictions (i.e. redefining predictor variables).

References

1. Statista. "Sports Betting Revenue in the U.S. 2018-2022." Statista, 7 Sept. 2023.

2. "The Psychology of Sports Betting." Virtual Behavioral Health with Kindbridge, 30 Sept. 2021.

3. "Paired Comparison Models with Tie Probabilities and Order Effects as a Function of Strength." Fields Institute for Research in Mathematical Sciences, 9 Jan. 2018.

4. Caley, Michael. "What Is the Best Method of Predicting Goals?" Cartilage Free Captain, 28 Feb. 2014.

5. MoneyPuck.com-about and How It Works.

6. Anzer, Gabriel, and Pascal Bauer. "A Goal Scoring Probability Model for Shots Based on Synchronized Positional and Event Data in Football (Soccer)." Frontiers in Sports and Active Living, vol. 3, 29 Mar. 2021, https://doi.org/10.3389/fspor.2021.624475.

7. Antipov, Evgeny A., and Elena B. Pokryshevskaya. "Interpretable Machine Learning for Demand Modeling with High-Dimensional Data Using Gradient Boosting Machines and Shapley Values." Journal of Revenue and Pricing Management, 14 Mar. 2020, https://doi.org/10.1057/s41272-020-00236-4.

8. "Index of Authors, Chairpersons and Organizers." 42nd IEEE International Conference on Decision and Control (IEEE Cat. No.03CH37475), 1 Jan. 2003, https://doi.org/10.1109/cdc.2003.1272917. Accessed 24 Apr. 2023.

9. Pieter Robberechts, and Jesse Davis. How Data Availability Affects the Ability to Learn Good XG Models. 14 Sept. 2020, pp. 17–27, https://doi.org/10.1007/978-3-030-64912-8_2. Accessed 18 May 2023.