Article

Construction of a Predictive Model for MLB Matches

Chia-Hao Chang

Department of Nursing, Chang Gung University of Science and Technology, Chiayi Campus, Chiayi 61363, Taiwan; howellchang@gmail.com; Tel.: +886-5-362-8800

Abstract: The main purpose of this article was to define a model that could defeat the online bookmakers’ odds, where the betting item considered was the first five innings of major league baseball (MLB) matches. The betting odds of online bookmakers have two purposes: first, they are used to quantify the amount of profit made by the bettors; second, they are regarded as a market equilibrium point between multiple bookmakers and bettors. If the bettors have a more accurate prediction model than the system used to produce betting odds, it will create a positive expected return for the bettors. In this article, we used the Markov process method and the runner advancement model to estimate the expected runs in an MLB match for the teams based on the batting lineup and the pitcher.

Keywords: betting odds; implied probability; Markov process; runner advancement model; expected return

1. Introduction

1.1. Background

Predicting patterns of behavior plays a pivotal role in many walks of life including the COVID-19 time series, animal movement, stock market system and the sports industry. However, a prediction’s quality is closely related to the forecasting fields involved. The inspiration for the main character of the movie “Moneyball”, Jeffrey Ma [1], mentioned in his book The House Advantage: Playing the Odds to Win Big in Business that it is much simpler to predict the unlisted cards in the hands of dealers than it is to predict the future return of the stock market because there are only 52 cards in a deck. The target of prediction discussed in this article is the probability of winning in baseball. The difficulty of predicting the probability of winning in baseball lies between that of predicting the unlisted cards and the stock market return. The probability of guessing the next card in the dealer’s hand depends on the cards and the number of cards that have already appeared on the table, while the prediction of the stock market return in the future includes the fundamental analysis factors and technical analysis indicators, as well as the political and economic factors at home and abroad. Fans of baseball may have heard that the pitcher’s impact on their games are greater than a hitter’s impact on their games. It is certain that pitchers are not the only factor affecting the outcome of baseball matches. As far as sports activities are concerned, being able to improve the outcome prediction accuracy of a match not only guides the trainer and sportsmen with the direction of their players’ management, but also creates lucrative opportunities in terms of sports wagering. At present, more than 120 countries around the world have issued sports lottery tickets. In Nevada alone, the so-called “city of gambling” in the United States, sports lottery sales in 2017 amounted to as much as 4.9 billion USD, a rise of 440% compared with 1984. In 2017, baseball-related bets accounted for 23% of all sports betting, second only to football [2]. In order to make a profit from sports wagering, many individuals and institutions have devoted themselves to the study of match predictions. However, it is difficult to improve the accuracy of predictions. Among the many sources of information that can be used to predict match outcomes, the betting odds offered by bookmakers are one of
the easiest sources from which the public can obtain professional predictions. The public can now easily obtain betting odds from online bookmakers. In order to make an ideal profit, bookmakers need to improve the outcome prediction accuracy of the game; therefore, they have developed several models that calculate the possibility of predicting the possible results in various sports competitions [3,4]. At present, there are three major betting odds systems: European, British and American odds. Most online bookmakers offer the function of switching the display of different betting odds in order to satisfy the betters’ preferences. The betting odds used in this article are European odds, also known as decimal odds, which are calculated as the reciprocal of the probability of an outcome, and can be explained with the mathematical equation as follows: Assuming that the probability of outcome $i$ is $\pi_i$, the European odds can be expressed as $\frac{1}{\pi_i}$.

There are many sources of betting odds. The fixed odds of online bookmakers alone have many sources to choose from, such as Bet365 [5] and Betclic, among others. Each bookmaker announces different betting odds and when the odds vary widely, it implies that violations of market efficiency exist. In fact, if everyone was to now bet on just one of those outcomes then the money coming in would be skewed one way for the bookmaker. In response to this, the bookmaker will increase their margin on the popular line to discourage betting, and will reduce their margin on the less popular line to encourage betting. At this point, such hedging action has changed the implied probability of the bookmaker. These can be factors that change the implied probability [6]. However, the announced betting odds can inform a betting decision. Utilizing the situation where bookmakers are forced to increase or decrease their betting odds, Kaunitz et al. [7] proposed a strategy to beat football bookmakers with their own numbers. Instead of building a forecasting model to compete with bookmakers’ predictions, they exploited the probability information implicit in the odds publicly available in the marketplace to find bets with mispriced odds. Shin [6,8,9] proposed Shin probabilities based on the assumption that bookmakers quote odds that maximize their expected profit in the presence of uninformed bettors and a known proportion of “insider” traders.

1.2. Literature Review

The motivation of this article comes from the independent development of the implied probability with even higher accuracy. In the past, there have been a lot of studies on how to beat bookmakers’ betting odds. Some have built statistical models using Power scores, Elo ratings, Maher–Poisson approaches, Bayesian network, and pi-ratings [10–14]. The method we propose here is to use a Markov chain [15], based on the D’Esopo and Lefkowitz runner advancement model (RAM) [16], to calculate the expected number of runs scored (ENRS) and the probability of leading in the first five innings in a baseball game. An inning is the basic unit of play in baseball and a full game is typically scheduled for nine innings. We deal with a stochastic process which is characterized by the rule that only the current state of the process can influence the choice of the next state. It means the process has no memory of its previous states. Such a process is called a Markov process, after the prominent Russian mathematician Andrey Markov (1856–1922). In baseball models, the states are usually the various runners and outs situations. The Markov chain assumption means that we are not interested in how we arrived at a particular situation. The relevant literature on these theories over recent years includes Hirotsu [17] and Hirotsu and Bickel [18], and the latter found the best order to achieve more than 0.5 in probability of winning the game against other possible batting orders. Hirotsu and Wright [19] proposed the formulation for obtaining the optimal pinch-hitting strategy under the designated hitter rule. Fritz and Bukiet [20] extended the Markov chain model to introduce an objective criterion for selecting the major league baseball (MLB) most valuable player (MVP) award winners and Cy Young award winners. Smith [21] presented a Markov chain model for predicting the scores and the winning team of major league baseball (MLB) games. Chang [22] modified the runner advancement model and presented the
significant factors that impact the ENRS. Chang [23,24] measured the number of contributed runs per game for each player. In this paper, we predicted the "runs of the first five innings" of an MLB match using the Markov process method and the runner advancement model before the game. We used historical data of the batting lineup vs. the pitcher from the MLB official website to estimate the expected number of runs for the teams. We tested the efficiency of MLB betting markets by examining the ability of the following two kinds of probability:

1. The Bet365 [5] online bookmaker announces its first five innings betting odds at 10 p.m. China Standard Time. The odds are pre-game odds obtained before the start of the match. This information is relevant, since a starting pitcher in MLB usually rests four or five days after pitching a game before pitching another. Therefore, most MLB teams have five or six starting pitchers on their rosters. These pitchers, and the sequence in which they pitch, is known as the rotation. For the most part the starting lineups will stay the same after posted. The choice of the bookmaker probably does not have a high effect on the homogeneity of prices due to the transparency of odds among online bookmakers and high competition in the market. The betting odds already include the profits of the bookmaker. Therefore, the inverse odds cannot be regarded as the implied probability directly. In the next section, we will introduce how to adjust bookmakers’ profits using basic normalization (BN) based on the literature.

2. We calculated the ENRS (based on batter vs. pitcher career statistics) and the probability of leading in the first five innings, which is the new implied probability (NIP) proposed in this article.

In a betting market, forecasting models are judged in terms of their accuracy and profitability. Wunderlich and Memmert [25] presented the counterintuitive relationship between accuracy and profitability in probabilistic forecasts in relation to betting markets. They said that betting accuracy should not be treated as a valid measure of a forecasting model. This article will also evaluate these two kinds of probability from the two perspectives of accuracy and profitability: first, the ranked probability score (RPS) [26] is a measure of how similar two probability distributions are and is used as a way to evaluate the quality of a probabilistic prediction; second, the expected value (EV) is the measure of what a bettor can expect to win or lose per bet placed on the same odds. The structure of the remainder of this article is as follows: In Section 2, the calculation methods of BN and NIP are introduced; the data sources of betting odds and competition scoring are introduced in Section 3, together with an explanation of how to use RPS and EV to evaluate the three kinds of probability; in Section 4, the results of three kinds of probability are presented; and suggestions and a discussion is provided in Section 5.

2. Determining Outcome Probabilities from Betting Odds

2.1. Basic Normalization

Assume that $\sigma = (o_1, o_2, \cdots, o_n)$ is the betting odds for a game match with $n$ outcomes, where $n \geq 2$, and the inverse odds is set to $\pi = (\pi_1, \pi_2, \cdots, \pi_n)$, therefore $\pi_i = 1/o_i$ could be regarded as the occurrence intensity of an outcome but cannot be regarded as the implied probability directly as the bookmaker has set the total sum of $\pi_i$ to be greater than 1, which means $\Pi = \sum_{i=1}^{n} \pi_i > 1$; we can regard $\Pi - 1$ as the bookmaker’s profit. In order to normalize the $\pi_i$, we divide $\pi_i$ by $\Pi$ to obtain the $p_i = \pi_i/\Pi$. We call $p_i$ the implied probability subject to basic normalization.

2.2. New Implied Probability (NIP)

Unlike BN, the NIP is not derived from the bookmaker’s betting odds, but uses Markov chain and RAM to calculate the ENRS in the first five innings. To calculate the NIP we first define a matrix $U_n$ containing the situations of scoring from half an inning, being on base, and outs of the current match, where the columns of $U_n$ represent the current scoring (there are 21 columns (0, 1, ..., 20) in total); the reason that the last line was set for
20 points was that the Boston Red Sox and Detroit Tigers once scored 19 points in 1953 among MLB’s highest scoring records in a single inning. The rows of \( U_n \) represent the current number of outs and base states. According to the framework of transition matrix required by the Markov process, 25 states can be summarized in Table 1 below [16] when considering the outs and base states faced by the offensive matchup, of which eight different base states are 0 out, 1 out, and 2 outs, that is, the states of runner on first, runner on second, runner on third, runners on first and second, runners on second and third, runners on first and third, runners on first, second, and third, and finally the state of 3 outs.

\( n \) means the \( n \)th batter is entering the game. As Bukiet’s [16] transition matrix \( T_{25 \times 25} \) is the basis of the Markov chain, we simplify the scoring effect generated by \( T_{25 \times 25} \) into the five categories of \( T_0, T_1, T_2, T_3, \) and \( T_4 \). For example, \( T_{(\cdot)}0 \) represents the probability matrix of scoring 0 points for the hitting outcome generated by the \( n \)th batter in the inning. Likewise, \( T_{(\cdot)}4 \) represents the probability matrix of scoring 4 points for the hitting outcome generated by the \( n \)th batter in the inning. Therefore, \( U_n \) provides information on the scoring outcome and the probability generated by the \( n \)th batter during the inning. As a result, before the start of the game \( U_{1\cdot} = 1 \), and the other matrix elements are all 0, which means that the occurrence probability of scoring 0 points, 0 runners on base, and 0 outs is 1. For readers to understand the \( U_n \) matrix more easily, the following lists the \( U_{1\cdot25} \) matrix after a first batter completed their hitting in the first inning, where the vector probabilities of the first column (scoring 0 point) and the second column (scoring 1 point) are as follows:

\[
U_{1\cdot1} = U_{0\cdot1} T_{(1)} 0
\]

and

\[
U_{1\cdot2} = U_{0\cdot2} T_{(1)} 0 + U_{0\cdot1} T_{(1)} 1
\]

The other vectors are all zero vectors. RAM simplifies the runner advancement and scoring outcomes into six cases, as shown in Table 2, with which they can be utilized to estimate the effect of runner advancement. Therefore, the probabilities of the compositional element of the \( U_{1\cdot} \) matrix for the first batter in the first inning are expressed as follows:

\[
U_{1\cdot2} = p(BB) + p(1B)
\]

\[
U_{1\cdot3} = p(2B)
\]

\[
U_{1\cdot4} = p(3B)
\]

\[
U_{1\cdot9} = p(Out)
\]

\[
U_{1\cdot2} = p(HR)
\]

The expected scoring of the first batter in the first inning is:

\[
0 \times \text{the sum of the probabilities of the 1st column} \\
+1 \times \text{the sum of the probabilities of the 2nd column} \\
+2 \times \text{the sum of the probabilities of the 21st column} \\
+20 \times \text{the sum of the probabilities of the 20th column}
\]

\( = 1 \times p(HR) \),

Therefore, the expected scores contributed by all the batters within a half-inning are as follows:

\[
0 \times U_{0\cdot1} T_{(2)} \cdots T_{(9)} T_{(1)} T_{(2)} \cdots \text{’s sum of the probabilities of the 1st column}
\]
\[ 1 \times U_{0}T_{1}T_{2} \cdots T_{(9)}T_{1}T_{2} \cdots \text{’s sum of the probabilities of the 2nd column +} \]
\[ \cdots \]
\[ 20 \times U_{0}T_{1}T_{2} \cdots T_{(9)}T_{1}T_{2} \cdots \text{’s sum of the probabilities of the 21st column,} \]

Following the rule until \( U_{n[k]} \) ’s sum of probabilities is greater than 0.999, that is, given the 21 states of scoring, if the marginal probability of the three outs is greater than 0.999, the half inning is immediately discontinued. Following this to precede the match for the first five innings, the expected scores at the end of the first five innings can be subsequently obtained.

Next, we explain how to use the \( U_{n[k]} \) at the end of the first five innings, namely, the scoring distribution of each matchup to calculate the leading, tied or behind probabilities of the game match. By adjusting a nine-inning winning expression proposed by Bukiet [16], we can obtain the leading probability as follows:

\[
NIP_{L} = \sum_{i=0}^{20} f(x = i) \sum_{j=0}^{i-1} f(y = j) \tag{5}
\]

While the behind probability is determined as follows:

\[
NIP_{b} = \sum_{i=0}^{19} f(x = i) \sum_{j=i+1}^{20} f(y = j) \tag{6}
\]

where \( f(x = i) \) represents the probability of matchup X scoring \( i \) points at the end of the first five innings, and \( f(y = j) \) represents the odds that matchup Y scoring \( j \) points at the end of the first five innings. In addition, Dolinar [27] proposed that \( f(x = i) \) and \( f(y = j) \) can be estimated using the negative binomial distribution (NBD), where the success event of the NBD is defined as the 15 outs and the failure event of NBD is defined as scoring \( i \) points. However, as the BN used in this article does not consider the tied probability, to compare on the same basis we used BN to define the NIP as follows:

\[
NIP = \frac{NIP_{L}}{NIP_{L} + NIP_{b}} \tag{7}
\]

### Table 1. The twenty-five codes for outs and base states.

| Code Number | (Base State, Number of Outs) | Code Number | (Base State, Number of Outs) | Code Number | (Base State, Number of Outs) | Code Number | (Base State, Number of Outs) |
|-------------|-----------------------------|-------------|-----------------------------|-------------|-----------------------------|-------------|-----------------------------|
| 1           | (0, 0)                      | 9           | (0, 1)                      | 17          | (0, 2)                      | 22          | (123, 0)                   |
| 2           | (1, 0)                      | 10          | (1, 1)                      | 18          | (1, 2)                      | 23          | (123, 1)                   |
| 3           | (2, 0)                      | 11          | (2, 1)                      | 19          | (2, 2)                      | 24          | (123, 2)                   |
| 4           | (3, 0)                      | 12          | (3, 1)                      | 20          | (3, 2)                      | 25          | (X, 3)                     |
| 5           | (12, 0)                     | 13          | (12, 1)                     | 21          | (12, 2)                     |             |                             |
| 6           | (13, 0)                     | 14          | (13, 1)                     | 22          | (13, 2)                     |             |                             |
| 7           | (23, 0)                     | 15          | (23, 1)                     | 23          | (23, 2)                     |             |                             |
| 8           | (123, 0)                    | 16          | (123, 1)                    | 24          | (123, 2)                    |             |                             |

□ □ □ □
Table 2. Runner advancement model of D’Esopo and Lefkowitz.

| Hitting Conditions | Outcome |
|--------------------|---------|
| Base Balled (BB)   | Batter safely reaches first base and does not advance unless there is a bases loaded. |
| One Base Hit (1B)  | Batter safely reaches first base; first base runner safely reaches second base, and the rest runners score points. |
| Two Base Hit (2B)  | Batter safely reaches second base; first base runner safely reaches third base, and the rest runners score points. |
| Three Base Hit (3B)| Batter safely reaches third base, and the rest runners score points. |
| Home Run (HR)      | Batter scores a point, and the rest runners score points. |

3. Data Processing

3.1. Research Tools

The data source was the 70 MLB matches with most batter vs. pitcher matchup stats in 2018 selected from the MLB statistics starting form 15 September 2018 to 30 September 2018. Betting odds were taken from the one of the first five innings of MLB matches announced by the Bet365 [5] online bookmaker. The starting pitchers and batting order of a match were taken from the MLB website [28], and the batter vs. pitcher matchup stats were taken from [29]. Among the indicators considered in RAM, we used the following formulas to calculate the different probabilities:

\[
BB\% = \frac{BB}{BB + 1B + 2B + 3B + HR + Out} \quad (8)
\]

\[
1B\% = \frac{1B}{BB + 1B + 2B + 3B + HR + Out} \quad (9)
\]

\[
2B\% = \frac{2B}{BB + 1B + 2B + 3B + HR + Out} \quad (10)
\]

\[
3B\% = \frac{3B}{BB + 1B + 2B + 3B + HR + Out} \quad (11)
\]

\[
HR\% = \frac{HR}{BB + 1B + 2B + 3B + HR + Out} \quad (12)
\]

\[
Out\% = \frac{Out}{BB + 1B + 2B + 3B + HR + Out} \quad (13)
\]

3.2. Rank Probability Score (RPS)

Epstein [26] proposed that the RPS should be used to evaluate the difference between the prediction probability and the real outcome. Constantinou and Fenton [30] said that the RPS is an agreed scoring rule to determine a forecasting model’s accuracy. In fact, RPS is a formula for calculating the linear distance between the prediction probability and the real outcome, so RPS is always greater than or equal to 0, and the closer RPS is to 0, the more accurate the prediction is. If we take the betting item discussed in this article as an example, we compare BN and NIP at the end of the first five innings with the real outcomes and evaluate their differences. If we consider the total accuracy of the betting on n matches, the calculation formulas are as follows:

\[
RPS_{BN} = \sum_{i=1}^{n} (BN_{ij} - 1_{ij})^2 \quad (14)
\]

\[
RPS_{NIP} = \sum_{i=1}^{n} (NIP_{ij} - 1_{ij})^2 \quad (15)
\]

where \(1_{ij}\) represents match \(i\)’s actual scoring leading indicator function.

3.3. Expected Value (EV)

The expected value (EV) is used to measure whether the betting odds for an outcome is a valuable betting opportunity. In short, if an implied probability multiplied by a net profit is greater than 1 after subtracting the implied probability multiplied by the bet cost,
then the bet is considered valuable. It can also be said that the larger the value is, the
greater the expected value on the betting odds under the implied probability. As far as
BN is concerned, because the bookmaker has reduced the betting odds in order to create
its own profits, the EV has to be negative, which means betting on the basis of BN is not
an appropriate strategy. Therefore, regarding the comparison of EV, we present only the
EV subject to the NIP. Thus, assuming that \( o_j = (o_{ij}, o_{ib}) \) is the betting odds for two out-
comes (leading and behind) for match \( i \), we obtain:

\[
EV_{NIP} = \sum_{i=1}^{n} [NIP_{ij} \cdot (o_{ij}-1) - NIP_{ib} \cdot (o_{ib}-1) - NIP_{ij}] \cdot 1_{ij} \quad (16)
\]

where \( n \) represents the total matches for betting, and besides calculating different implied
probabilities, the above two equations are consistent; \( 1_{ij} \) represents the real leading indica-
tor function of match \( i \); and \( 1_{ib} \) represents the real behind indicator function of match \( i \).

4. Results

4.1. Comparison of the Prediction Probability and the Real Outcomes

We compared the accuracy of the three prediction probabilities (Table A1). The leading
probability in Table A1 refers to the team in the left column; for example, for the match
listed in the first row, this refers to the match held on 15 September 2018, where the match-
ups were between New York Yankees (NYY) and Toronto Blue Jays (TOR). NYY’s first
five-inning leading probabilities subject to the prediction of BN, NIP and NIP–NBD were
0.71, 0.83, and 0.84 respectively. The values presented in the last row are the real match
outcomes, where 1 indicates that the team in the left field (NYY) leads, 0 indicates that the
team in the right field (TOR) leads, and “-” indicates a tie. Tables 3 and 4 present the RPS
and EV results for a specific date (15 September 2018) and team (NYY), respectively. The
average RPS over 70 matches was \( RPS_{BN} = 0.24, \ RPP_{NIP} = 0.17, \ RPS_{NIP-NBD} = 0.15, \) and
a Wilcoxon signed-rank test revealed that these differences were statistically significant (\( p = 0.02 \)).

4.2. Comparison of the Expected Value

Without considering the tied situations (a total of 60 matches), we compared the pre-
diction accuracy of NIP and NIP–negative binomial distribution (NIP–NBD) according to
the information provided in Table A1, which was \( NIP = 45/60 = 75\% , \) and
\( NIP \_NBD = 46/60 = 76.7\% . \) The NIP–NBD exhibited higher consistency with the real
outcomes. In terms of the expected value, the betting odds \( (o_1, o_2) \) for matches in the
second row were 1.6 and 2.35, respectively, while for the five-inning outcome the party
with betting odds of 2.35 leads, and the expected values calculated according to NIP and
NIP–NBD were \(-0.44\) and \(-0.48\), respectively. The expected values using the predictions
for 60 matches were \( EV_{NIP} = 10.15, \) and \( EV_{NIP-NBD} = 11.51, \) which means if one unit bet
was cast on each match, the 60-match expected values subject to NIP and NIP–NBD were
10.15, and 11.51, respectively.

Table 3. Comparison of rank probability scores (RPSs) and expected values (EVs) subject to basic normalization (BN), new
implied probability (NIP), and NIP–negative binomial distribution (NIP–NBD) for 15 September 2018.

| Date          | Matches | RPS\(_{BN}\) | RPS\(_{NIP}\) | RPS\(_{NIP-NBD}\) | \( O_1 \) | \( O_2 \) | Lead | EV\(_{NIP}\) | EV\(_{NIP-NBD}\) |
|---------------|---------|-------------|-------------|------------------|---------|---------|------|------------|------------------|
| 15 September 2018 | NYY TOR  | 0.08        | 0.03        | 0.02             | 1.33    | 3.3     | 1    | 0.1        | 0.12              |
| 15 September 2018 | WSH ATL  | 0.35        | 0.58        | 0.6              | 1.6     | 2.35    | 2    | -0.44      | -0.48             |
| 15 September 2018 | MIA PHI  | 0.15        | 0.04        | 0.04             | 2.45    | 1.57    | 2    | 0.24       | 0.26              |
| 15 September 2018 | SF COL   | 0.3         | 0.53        | 0.59             | 2.1     | 1.71    | 1    | -0.43      | -0.51             |
| 15 September 2018 | DET CLE  | 0.4         | 0.22        | 0.22             | 2.6     | 1.52    | 1    | 0.37       | 0.38              |
| 15 September 2018 | MIN KC   | 0.18        | 0.25        | 0.32             | 1.66    | 2.2     | 1    | -0.17      | -0.28             |
Table 4. Comparison of RPSs and EVs subject to BN, NIP and NIP–NBD for team New York Yankees (NYY).

| Date             | Matches | RPS<sub>BN</sub> | RPS<sub>NIP</sub> | RPS<sub>NIP–NBD</sub> | O<sub>1</sub> | O<sub>2</sub> | Lead | EV<sub>NIP</sub> | EV<sub>NIP–NBD</sub> |
|------------------|---------|------------------|------------------|---------------------|-----------|-----------|------|----------------|------------------|
| 15 September 2018| NYY TOR | 0.08             | 0.03             | 0.02                | 1.33      | 3.3       | 1    | 0.1           | 0.12             |
| 19 September 2018| BOS NYY | 0.33             | 0.32             | 0.35                | 2.25      | 1.64      | 1    | –0.03         | –0.07            |
| 21 September 2018| BOS NYY | 0.16             | 0.13             | 0.13                | 2.4       | 1.58      | 2    | 0             | 0.02             |
| 22 September 2018| NYY BAL | 0.07             | 0.04             | 0.03                | 1.3       | 3.5       | 1    | 0.04          | 0.07             |
| 27 September 2018| TB NYY  | 0.38             | 0.94             | 0.94                | 2.5       | 1.55      | 1    | –0.92         | –0.92            |
| 29 September 2018| NYY BOS | 0.19             | 0.02             | 0.02                | 1.68      | 2.15      | 1    | 0.43          | 0.46             |

5. Discussion and Application

In this article, we introduced the most accessible information of betting odds for the general public, which are the betting odds announced by bookmakers. Considering the profits of the bookmakers, we use BN to restore the betting odds information. Markov chain and RAM are important theoretical foundations for predicting baseball scoring. In this article, we used these models to predict MLB match scoring and to calculate the implied probability NIP. By evaluating the probability values, including prediction accuracy and the evaluation of expected values, we proved that NIP has its advantages in terms of the number of matches (n = 70) considered, whether in terms of RPS or EV. In fact, during the theoretical analysis for the 70 matches in this article, if we restore the very moment of betting, where the outcomes were unknown, the total return for the 70 matches according to the prediction probability models of NIP, and NIP–NBD are 23.89 and 22.69, respectively, converted to return on investment (ROI) as 34.13% and 32.41%.

By its very nature the sport of baseball is highly suited to adaption as a Markov chain as by its very nature it is split into discrete standalone plays [31–33]. Through this paper we have been able to derive a first five-inning scores for a match. The transition probabilities, which derived on the hitting condition are used for our modeling of baseball as a Markov process. The input factors to our model is the roster of the home team, the roster of the away team, statistics of all players on each team leading up to a game. Baseball is unique in how much data is available online. Statistics on every plate appearance in MLB since 1921 are available for free. Bookmakers can use this abundance of data to improve odds.

NIPs were not derived from the bookmaker’s betting odds but used batter/pitcher matchup history stats between any pair of players. Fans of baseball games may have heard that the pitcher’s impact on a game is greater than a hitter’s impact. A team manager usually wants to get 5–6 innings from his starter pitchers. Sometimes, despite pitchers having good arms, good quality pitches and high throwing velocity, they do not have the stamina for those 5–6 innings. Moreover, a starting pitcher must pitch at least five innings to qualify for the win. In summary, a first five innings bet allows us to focus on a much smaller range of factors when searching for value in wagers.

As the first five-inning betting market is small, there are not many historical data sources that can be obtained directly. Therefore, the 70-match betting odds and the historical batter vs. pitcher stats and starting batting order referred to in this article are difficult for general researchers to obtain directly through computerization. Therefore, it is a limitation of this article that in practical applications, the relevant information, such as batting order and betting odds, need to be input before betting, so it takes a lot of time to obtain the NIP. In addition, compared with similar articles, the number of samples in this study was low. In future studies, we would increase the number of matches considered.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.
Data Availability Statement: Data available in a publicly accessible repository that does not issue DOIs. Publicly available datasets were analyzed in this study. This data can be found here: www.mlb.com.

Conflicts of Interest: The author declares no conflict of interest.

Appendix A

Table A1. Prediction probabilities subject to BN, NIP and NIP–NBD, and real outcomes.

| Date                | Matches | BN   | NIP  | NIP–NBD | Results |
|---------------------|---------|------|------|---------|---------|
| 15 September 2018   | NYY TOR | 0.71 | 0.83 | 0.84    | 1       |
| 15 September 2018   | WSH ATL | 0.59 | 0.76 | 0.78    | 0       |
| 15 September 2018   | MIA PHI | 0.39 | 0.21 | 0.20    | 0       |
| 15 September 2018   | SF COL  | 0.45 | 0.27 | 0.23    | 1       |
| 15 September 2018   | DET CLE | 0.37 | 0.53 | 0.53    | 1       |
| 15 September 2018   | MIN KC  | 0.57 | 0.50 | 0.43    | 1       |
| 16 September 2018   | MIN KC  | 0.47 | 0.11 | 0.10    | 0       |
| 16 September 2018   | CIN CHC | 0.40 | 0.74 | 0.93    | -       |
| 16 September 2018   | ARI HOU | 0.36 | 0.15 | 0.12    | 0       |
| 16 September 2018   | COL SF  | 0.56 | 0.36 | 0.35    | 0       |
| 18 September 2018   | ATL STL | 0.54 | 0.13 | 0.11    | 0       |
| 18 September 2018   | ARI CHC | 0.58 | 0.53 | -       | -       |
| 19 September 2018   | BOS NYY | 0.42 | 0.43 | 0.41    | 1       |
| 19 September 2018   | NYM PHI | 0.36 | 0.59 | 0.59    | 1       |
| 19 September 2018   | CWS CLE | 0.31 | 0.41 | 0.34    | 0       |
| 19 September 2018   | CIN MIL | 0.36 | 0.72 | 0.74    | 1       |
| 20 September 2018   | CIN MIL | 0.39 | 0.12 | 0.11    | 0       |
| 20 September 2018   | NYM PHI | 0.54 | 0.36 | 0.35    | 0       |
| 21 September 2018   | CWS CLE | 0.36 | 0.60 | 0.61    | 1       |
| 21 September 2018   | BOS NYY | 0.40 | 0.36 | 0.36    | 0       |
| 21 September 2018   | OAK ANA | 0.54 | 0.88 | 0.93    | 1       |
| 22 September 2018   | COL ARI | 0.47 | 0.26 | 0.22    | 0       |
| 22 September 2018   | HOU ANA | 0.68 | 0.53 | 0.53    | 1       |
| 22 September 2018   | SD LAD  | 0.29 | 0.95 | 0.99    | 1       |
| 22 September 2018   | NYY BAL  | 0.73 | 0.80 | 0.82    | 1       |
| 22 September 2018   | PIT MIL  | 0.43 | 0.65 | 0.75    | 1       |
| 23 September 2018   | COL ARI | 0.38 | 0.84 | 0.86    | 1       |
| 23 September 2018   | ATL PHI  | 0.58 | 0.66 | 0.67    | 1       |
| 23 September 2018   | HOU ANA | 0.71 | 0.64 | 0.67    | 1       |
| 23 September 2018   | SEA TEX  | 0.50 | 0.87 | 0.89    | 1       |
| 24 September 2018   | COL ARI | 0.46 | 0.48 | 0.48    | 1       |
| 24 September 2018   | OAK MIN | 0.64 | 0.67 | 0.70    | 0       |
| 24 September 2018   | TOR TB  | 0.38 | 0.15 | -       | 0       |
| 25 September 2018   | ARI LAD  | 0.37 | 0.42 | 0.41    | 1       |
| 25 September 2018   | BAL BOS  | 0.31 | 0.07 | 0.05    | 0       |
| 25 September 2018   | CWS CLE  | 0.33 | 0.45 | 0.35    | -       |
| 25 September 2018   | PIT CHC  | 0.41 | 0.06 | 0.03    | 1       |
| 25 September 2018   | OAK SEA  | 0.41 | 0.24 | 0.23    | 1       |
| 25 September 2018   | SD SF    | 0.44 | 0.58 | 0.59    | 1       |
| 25 September 2018   | HOU TOR  | 0.64 | 0.60 | 0.59    | 1       |
| 26 September 2018   | ANA TEX  | 0.59 | 0.88 | -       | 0       |
| 26 September 2018   | CLE CWS  | 0.65 | 0.71 | 0.72    | 1       |
| Date              | Team 1 | Team 2 | Probability | Win Probability | Draw Probability | Loss Probability |
|-------------------|--------|--------|-------------|-----------------|------------------|-----------------|
| 26 September 2018 | OAK    | SEA    | 0.50        | 0.87            | 0.90             | 1               |
| 26 September 2018 | SF     | SD     | 0.50        | 0.49            | 0.48             | 0               |
| 27 September 2018 | ARI    | LAD    | 0.49        | 0.98            | 0.95             | 1               |
| 27 September 2018 | BOS    | BAL    | 0.77        | 0.98            | 1.00             | -               |
| 27 September 2018 | CWS    | CLE    | 0.38        | 0.07            | 0.03             | 0               |
| 27 September 2018 | CHC    | PIT    | 0.63        | 0.36            | 0.35             | 1               |
| 27 September 2018 | COL    | PHI    | 0.65        | 0.97            | 0.98             | 1               |
| 27 September 2018 | MIN    | DET    | 0.57        | 0.54            | 0.54             | 1               |
| 27 September 2018 | NYM    | ATL    | 0.62        | 0.59            | 0.58             | -               |
| 27 September 2018 | OAK    | SEA    | 0.50        | 0.47            | 0.47             | 1               |
| 27 September 2018 | MIL    | STL    | 0.51        | 0.61            | 0.63             | 1               |
| 27 September 2018 | TB     | NYY    | 0.38        | 0.03            | 0.03             | 1               |
| 28 September 2018 | CHC    | PIT    | 0.61        | 0.89            | 0.91             | 1               |
| 28 September 2018 | NYM    | ATL    | 0.48        | 0.15            | 0.11             | 1               |
| 29 September 2018 | NYY    | BOS    | 0.56        | 0.85            | 0.87             | 1               |
| 29 September 2018 | PIT    | CIN    | 0.47        | 0.55            | 0.55             | 1               |
| 29 September 2018 | KC     | CLE    | 0.38        | 0.38            | 0.33             | 0               |
| 29 September 2018 | MIA    | NYM    | 0.44        | 0.71            | 0.75             | 1               |
| 29 September 2018 | SD     | ARI    | 0.41        | 0.10            | 0.06             | -               |
| 29 September 2018 | SEA    | TEX    | 0.63        | 0.71            | 0.73             | 1               |
| 29 September 2018 | LAD    | SF     | 0.63        | 0.63            | 0.65             | 1               |
| 29 September 2018 | TB     | TOR    | 0.62        | 0.33            | 0.31             | 1               |
| 30 September 2018 | COL    | WSH    | 0.51        | 0.07            | 0.06             | 0               |
| 30 September 2018 | KC     | CLE    | 0.30        | 0.31            | 0.30             | -               |
| 30 September 2018 | CWS    | MIN    | 0.44        | 0.02            | 0.04             | 0               |
| 30 September 2018 | MIA    | NYM    | 0.40        | 0.51            | 0.50             | -               |
| 30 September 2018 | PHI    | ATL    | 0.59        | 0.64            | 0.67             | -               |
| 30 September 2018 | TOR    | RAY    | 0.30        | 0.03            | 0.02             | -               |

References

1. Jeffrey, M. *The House Advantage: Playing the Odds to Win Big in Business*, 1st ed.; St. Martin's Press: New York, NY, USA, 2010.
2. Street & Smith's Sports Business Journal. Available online: https://www.sportsbusinessdaily.com/Journal/Issues/2018/04/16/World-Congress-of-Sports/Research.aspx (accessed on 16 April 2018).
3. Cantinotti, M.; Ladouceur, R.; Jacques, C. Sports betting: Can gamblers beat randomness? *Psychol. Addict. Behav.* 2004, 18, 143–147.
4. Garcia, J.; Perez, L.; Rodriguez, P. Football pools sales: How important is a football club in the top divisions? *Int. J. Sport Financ.* 2008, 3, 167–176.
5. Bet365. Available online: https://help.bet365.com/product-help/sports/rules/baseball (accessed on 30 December 2020).
6. Shin, H.S. Measuring the incidence of insider trading in a market for state-contingent claims. *Econ. J.* 1993, 103, 1141–1153.
7. Beating the Bookies with Their Own Numbers—and How the Online Sports Betting Market is Rigged. Available online: https://arxiv.org/abs/1710.02824 (accessed on 30 December 2020).
8. Shin, H.S. Optimal betting odds against insider traders. *Econ. J.* 1991, 101, 1179–1185.
9. Shin, H.S. Prices of state contingent claims with insider traders, and the favourite-longshot bias. *Econ. J.* 1992, 102, 426–435.
10. Dixon, M.; Coles, S. Modeling association football scores and inefficiencies in the football betting market. *Appl. Stat.* 1997, 46, 265–280.
11. Maher, M.J. Modeling association football scores. *Stat. Neerl.* 1982, 36, 3.
12. Vlastakis, N.; Dotsis, G.; Markellos, R.N. Nonlinear modelling of European football scores using support vector machines. *Appl. Econ.* 2008, 40, 111–118.
13. Constantinou, A.C.; Fenton, N.E. Profiting from arbitrage and odds biases of the European football gambling market. *J. Gambbl. Bus. Econ.* 2013, 7, 41–70.
14. Constantinou, A.C.; Fenton, N.E.; Martin Neil, M. Profiting from an inefficient association football gambling market: Prediction, risk and uncertainty using Bayesian networks. *Knowl. Based Syst.* 2013, 50, 60–86.
15. Bukiet, B.; Harold, E.R.; Palacios, J.L. A Markov chain approach to baseball. *Operat. Res.* 1997, 45, 14–23.
16. D’Esopo, D.A.; Lefkowitz, B. The distribution of runs in the game of baseball. In *Optimal Strategies in Sports*; Ladany, P., Machol, R.E., Eds.; North-Holland Publishing Company: Amsterdam, Holland, 1997.
17. Hirotsu, N. Reconsideration of the best batting order in baseball: Is the order to maximize the expected number of runs really the best? *J. Quant. Anal. Sports* 2011, 7, 1–12.
18. Hirotsu, N.; Bickel, J.E. Optimal batting orders in run-limit-rule baseball: A Markov chain approach. *IMA J. Manag. Math.* 2014, 27, 297–313.
19. Hirotsu, N.; Wright, M. Modeling a baseball game to optimize pitcher substitution strategies using dynamic programming. *Econ. Manag. Optim. Sports* 2004, 62, 131–161.
20. Fritz, K.; Bukiet, B. Objective method for determining the most valuable player in major league baseball. *Int. J. Perform. Anal. Sport* 2010, 10, 152–169.
21. Smith, Z.J. A Markov Chain Model for Predicting Major League Baseball. Ph.D. Thesis, University of Texas, Austin, TX, USA, 2016.
22. Chang, C.C. Runner advancement model application in CPBL. *J. Taiwan Intell. Technol. Appl. Stat.* 2017, 15, 31–46.
23. Chang, C.C. Using new sabermetrics index to provide CPBL hitter salaries. *J. Taiwan Intell. Technol. Appl. Stat.* 2018, 16, 19–30.
24. Chang, C.C. Improving Taiwanese baseball data analysis goes hand in hand with U.S.A. *Sci. Educ. Monthly* 2018, 407, 2–18.
25. Wunderlich, F.; Memmert, D. Are betting returns a useful measure of accuracy in (sports) forecasting? *Int. J. Forecast.* 2020, 36, 713–722.
26. Epstein, E. A scoring system for probability forecasts of ranked categories. *J. Appl. Meteorol.* 1969, 8, 985–987.
27. STATS.SEANDOLINAR.com. Available online: https://stats.seandolinar.com/mlb-run-distribution-neg-binomial/ (accessed on 30 December 2020).
28. MLB. Available online: https://www.mlb.com/starting-lineups/ (accessed on 30 December 2020).
29. MLB. Available online: http://mlb.mlb.com/stats/sortable_batter_vs_pitcher.jsp?season=2018 (accessed on 30 December 2020).
30. Constantinou, A.C.; Fenton, N.E. Solving the problem of inadequate scoring rules for assessing probabilistic football forecast models. *J. Quant. Anal. Sports* 2012, 8, 1–14.
31. Albert, J. Streaky hitting in baseball. *J. Quant. Anal. Sports* 2008, 4, 1–32.
32. Rump, C. Data clustering for fitting parameters of a Markov chain model of multi-game playoff series. *J. Quant. Anal. Sports* 2008, 4, 1–19.
33. Cserepy, N.; Ostrow, R.; Weems, B. *Predicting the Final Score of Major League Baseball Games*; Stanford University: Stanford, CA, USA, 2015.