Diamonds on the Line: Profits through Investment Gaming

Clayton Graham
DePaul University
Chicago, Illinois, USA 60604
Email: thelastviking@aol.com

ABSTRACT

With over a trillion dollars[1] being risked on worldwide sports gambling every year, the interest to modeling game performance in general and baseball in particular has gained growing popularity. Integrating baseball game modeling with analytically based gambling, allows for these two elements to be exploited with a single objective: profiting from the marketplace inequities between the game (production) and betting markets (price and lines). Two questions will be addressed: First, can an accurate baseball gaming model be derived and used to calculate the probability of winning and the economic consequence predicated upon the betting line? Second, what is the optimal bet size based upon the risk tolerances (operational constraints) of the investor? Included will be the derivation of a production function which can be used to calculate the probability of a winning team. Defining the implication of the betting line will address cost, payoffs, and the implied probabilities of winning. Expected Return on Investment and Betting Edge will provide an economics perspective.

1 Introduction

Ever since Alexander Cartwright put his New York Knickerbockers Baseball Club on the field in 1845, there have been those on the sidelines wagering on the outcome of the day's contest. The objective of investment gaming is simply to maximize expected profits over a baseball season (or fraction thereof), subject to investor risk limitations. Baseball is unique in that any prediction of an outcome is a function of direct matchups (batter vs. pitcher), aggregate team performance, and the field of play. To place the analysis in perspective, the following steps will be taken:

- Build a predictive production function with resultant probability of winning,
- Define the betting line along with its implications,
- Establish an economics relationship between the production function and line,
- Create a risk-return based investment function compatible with the production model,
- Quantify results of the model.

The nomenclature used throughout the paper is detailed in Appendix A.

2 Production Function and the Probability of Winning

Building a reliable predictive Production Function requires a pragmatic selection of input and output variables. There are just two purposes for a batter: get on base and drive those on base around to score. Consequently, the inputs to scoring include the measures of singles, doubles, triples, home runs, and base on balls. Other situations may arise where a batter can get on base (hit by pitch, drop third strike by catcher, etc.) or move along the base path (stolen base). In a prediction mode they are little more than noise.

The eventual desired output is runs scored per game by each team. Getting to runs per game directly may lead to inaccuracies because the number of innings played varies from game to game. Figure 1 demonstrates the various scoring trends which can lead to ambiguities of interpretation. Since the expected number of outs for a predicted game is 27 (9 innings * 3 outs/inning), the key output metric to be used is runs per out. Hence, runs/out is comparable among and between road and home games and leagues.

[1] "Sports Betting Tops One Trillion US Dollars Says Bookmaker to the Billionaires", http://www.prnewswire.com/news-releases/sports-betting-tops-one-trillion-us-dollars-says-bookmaker-to-the-billionaires-273768381.html, accessed December 5, 2014.

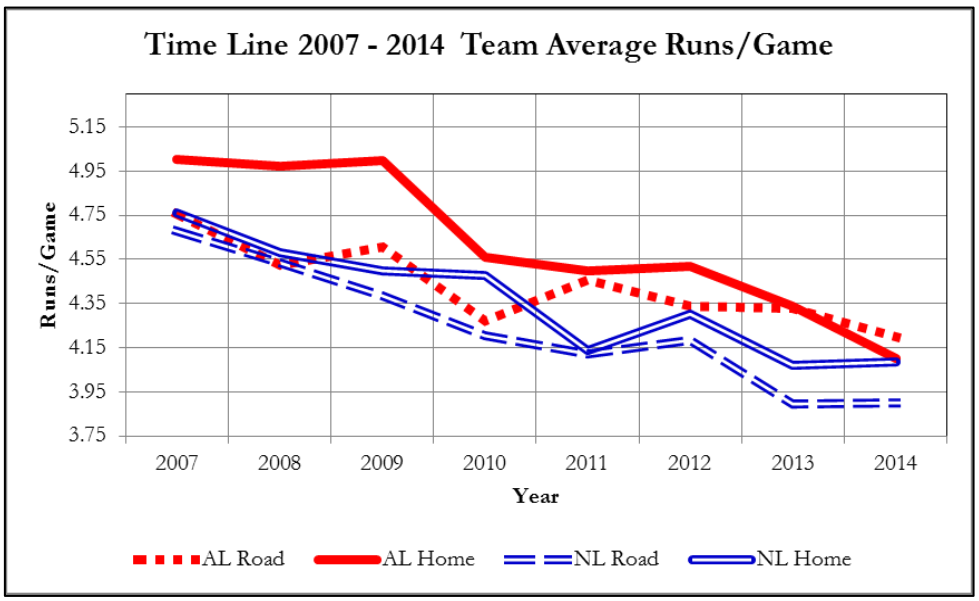


Figure 1 - Average Runs/Game/Team - Trending Downward
Source: http://mlb.mlb.com/stats/, accessed December 5, 2014.

The following is the overall forecasting regression equation:

Runs/Out = f(singles, doubles, triples, home runs, and base on balls)	**2.11**
Runs/Out = .3928 * %1 + .9053 * (%2+3) + 1.7361 * %HR + .1716 * %BB	**2.12[2]**
Predicted: EVR/9I = 27 * Runs/Out	**2.13**

In equation 2.12, percentage contribution by input variable is used in order to maintain accurate comparability between teams, individual and grouped players. A zero intercept is forced so as to assure credence to only actual factors of production.

"Pythagorean Theory of Baseball" was a metric put together by Bill James and simply stated is:

$$\textbf{Expected Winning Percentage} = \textbf{Runs Scored}^2 / (\textbf{Runs Scored}^2 + \textbf{Runs Allowed}^2) \quad \textbf{2.21}$$

In our nomenclature:

$$EW\% = RS^2 / (RS^2 + RA^2) \quad \textbf{2.22}$$

There are many iterations of this well-known formula (2.22). The exponents (originally 2 as above) for baseball have been recalculated ranging from 1.83 to over 2.0 [1]. For consistency and simplicity the exponent 2 will be used. The runs scored/allowed have generally been associated with a team's performance over a season. By estimating a single game's run scoring for each team, equation 2.22 can be used to calculate the probability of winning a single game. The density function of runs/9 innings is extrapolated from runs/out and is in Figure 2. To forecast runs it is essential to have a family of density functions that may be calculated from a team's (players') performance characteristics. Using the Gamma Function fulfills this requirement, see overlay in Figure 2. The Gamma Function not only fits the source data well[3] , it's parameters (α and β) may readily be derived with traditional statistics of mean and variance.

From the moment generating function, the first moment (μ) and second moments (δ^2) are:

$\mu = \alpha \beta$	**2.31**
$\delta^2 = \beta^2 \alpha$	**2.32[4]**

[2] See Appendix B for runs/out regression equation.
[3] See insert Figure 2.
[4] Appendix C contains team statistics used in parameter specification in modeling.

Solving above simultaneous equations (2.31 and 2.32) yields both α and β in terms of the readily calculated mean and variance:

$$\alpha = \mu^2 / \delta^2 \qquad 2.33$$
$$\beta = \delta^2 / \mu \qquad 2.34$$

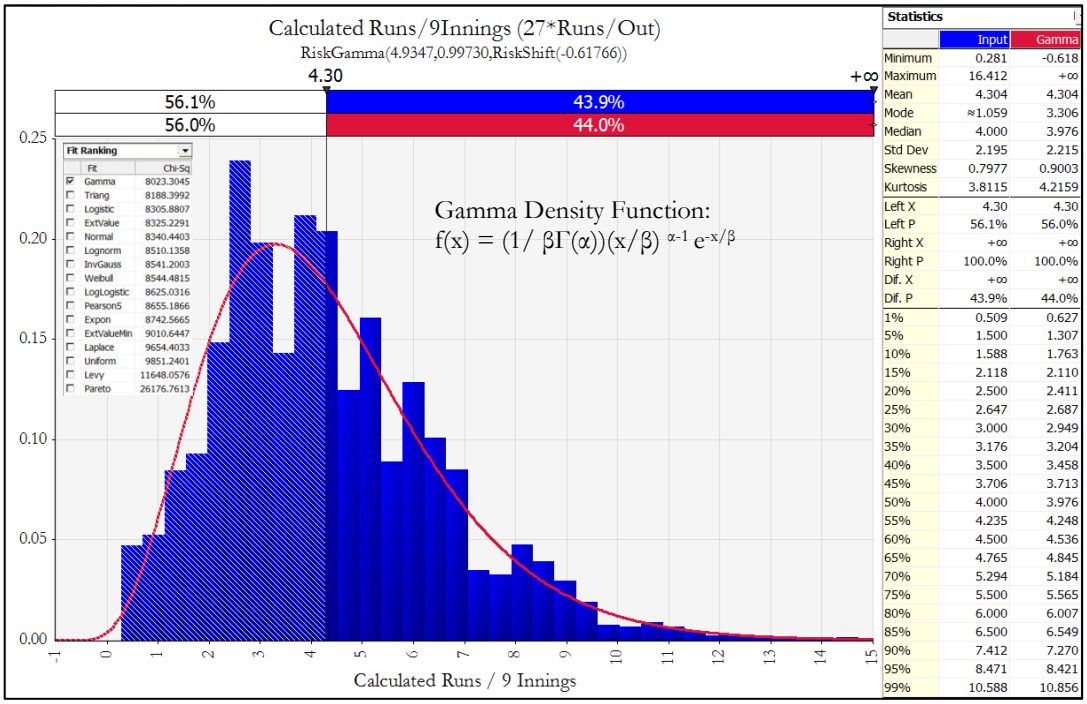


Figure 2 - Normalized Runs / Game / Team per 9 Innings 2011-13, Gamma Distribution Overlay Source: http://www.retrosheet.org/gamelogs/, accessed December 5, 2014.
Distribution fitting and resultant graphics created in Decision Tools Suite by Palisade Corporation.

Hence, the expected winning percentage per Equation 2.21 may now be written as a combination of density functions:

$$EW\% = (\Gamma(\alpha_{RS}, \beta_{RS}))^2 / ((\Gamma(\alpha_{RS}, \beta_{RS}))^2 + (\Gamma(\alpha_{RA}, \beta_{RA}))^2 \qquad 2.35$$

The prior formula can now be executed through a Monte Carlo simulation to accurately represent the distribution of winning percentages and winning margins. Figure 3 shows the results of the simulation. The distribution of scoring difference leads to a probability of the home team winning of 56% in contrast to the 62% calculated using equation 2.22. , This variation is not surprising considering the differences between the Gamma distributions for the Mariners and Tigers. The process represented in Equation 2.35 is the primary formula for quantifying the probability of winning. However the Gamma functions arguments need to be modified for the influence of the ball park in which the game is being played (park factor).

First is necessary to estimate expected run production for the day's batter/pitcher matchup. The detailed history of the starting pitcher and batters will be tabulated and summarized for each team. Appendix D contains results of the tabulations for the August 16th game between the Seattle Mariners and Detroit Tigers.

$$\%1 = 100 * \sum_{i=1}^{n}(1B_i) \ / \ PPA \quad \{PPA=productive\ plate\ appearance\} \qquad 2.41$$
$$\%2 + \%3 = 100 * \sum_{i=1}^{n}(2B_i + 3B_i) \ / \ PPA \qquad 2.42$$
$$\%HR = 100 * \sum_{i=1}^{n}(HR_i) \ / \ PPA \qquad 2.43$$
$$\%BB = 100 * \sum_{i=1}^{n}(BB_i) \ / \ PPA \qquad 2.44$$

For n batters for each team where relevant data is available

Inserting the results from equations 2.41-2.44 into Equation 2.12 and multiplying by 27 (runs/out/ game/club) will generate the preliminary estimate for a teams expected run production. Those results need to be adjusted further to account for the inherent bias represented by the difference in Park Factors between the road and home clubs.

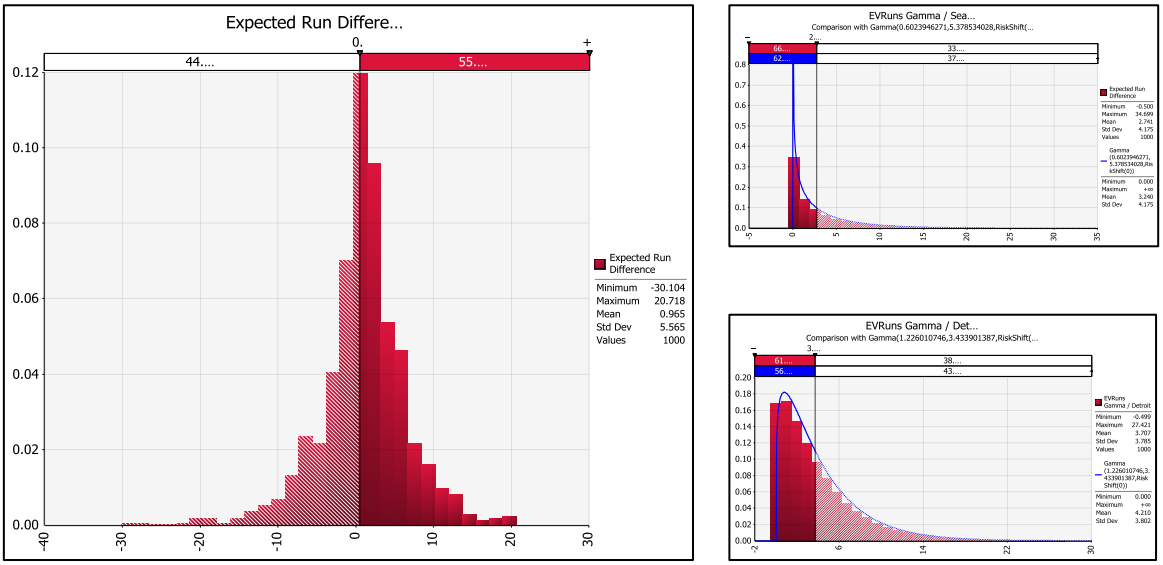


Figure 3 - Expected Winning Margin Simulated using Mariners' and Tiger's Derived Gamma Distributions through a Monte Carlo Simulation

The Park Factor is an index that measures the difference between runs scored in a team's home park and road games[5]. Figure 4 demonstrates the wide range of influence that the park itself has upon scoring. To equalize the influence of the park factors between the competing teams, expected value of run production needs to be scaled upward or downward. Since the starting pitcher and batters have likely played at both fields the park factor adjustment is simply the average of the two parks factors. In addition, the starting pitchers average about six innings per game[6]. The average runs/out during the last three innings is 91% of that tallied during the first six innings. Hence, the park factor adjustment and normalized expected runs/game becomes:

$$\textbf{park factor adjustment} = (\text{Average}(PF_{RD}, PF_{HM}) * (6*100 + 3*91)/9 \qquad 2.51$$

$$\textbf{normalized expected runs/game} = (EVR/9I) * \text{average}(PF_{RD}, PF_{HM}) * .97 \qquad 2.52$$

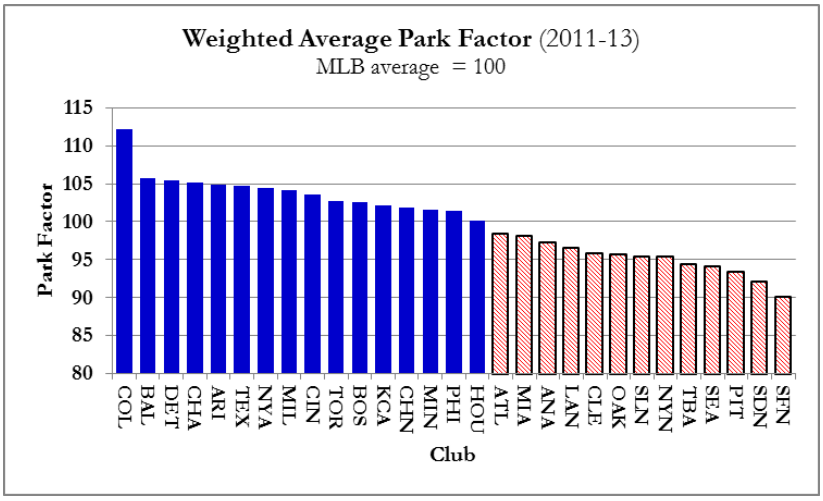


Figure 4 - Weighted Average of Park Factors (2011-2013) Shows Variation
Source: http://espn.go.com/mlb/stats/parkfactor/_/sort/HRFactor

At this juncture the production function meets the objective of predicting game scores. Stochastically calculating the probability of winning the game uses Equation 2.35 with Monte Carlo simulation. The next step is to examine the betting line.

[5] A detailed discussion on park factor (adjustment) can be found at: http://www.baseball-reference.com/about/parkadjust.shtml, accessed December 5, 2014.
[6] Tabulated from game logs 2011-2013, http://www.retrosheet.org/gamelogs/, accessed December 5, 2014.

3 Line, "Vig", Juice and other Mysteries

The betting line sets the cost of the bet, resulting payoff and establishes an implied probability of winning. Let's examine the following Table:

Table 1 - Example of Money Line Terms and Calculations[7]

August 16, 2014	Clubs	Money Line	@Risk Invest	Wager Pay-out	Implied Winning Probability IP(W)	Normalized Implied Winning Probability NIP(W)
Road Team	Seattle	-113	113	100	53.05%	52.1%
Home Team	Detroit	105	100	105	48.78%	47.9%
Totals or Aggregates			213	202.5	101.83%	100%
"Vig"			(213-202.5)/213 ≈ 5%			

The opening line is usually established when the first large Las Vegas Casino posts it. Setting the line is a result of research, contacts, experience, intuition[8] and divine guidance from the gnomes of Zurich.

Money line wagers are the most common form of betting on baseball. Just what does Table 1 mean? Let's examine some of the relevant elements from Table:

- Money Line (ML) Seattle -113 bet 113 (@ risk) to win 100 (wager)
  Detroit 105 bet 100 (@ risk) to win 105 (wager)

- Implied winning probability[9] (IP(W)): Seattle 113/(113 +100) = 53.05%
  Detroit 100 /(105 + 100) = 48.78%
  Total = 101.83%

- Normalized implied winning probability NIP(W):
  Seattle 53.05%/1.0183 = 52.1%
  Detroit 48.78% / 1.0183 = 47.90%
  Total = 100.0%

- Vigorous ("vig" or juice) house profit: Bets @ risk received: 113 + 100 = 213
  Payout wager: (100 or 105) = 202.5
  Difference = 10.5
  House return on bets @ risk 10.5/213 ≈ 5%

Integrating the money line's cost, payouts, and implied probabilities of winning together with the production functions expected scoring and the game's predicted probabilities of winning yields the economics consequences.

4 Economics

There are two principle methods to examine the economic outcome of a sports gaming investment:

First is the Expected Value of Return on Investment and is defined as:

$$\mathbf{EVROI_{HM} = [(P(W_{HM}) * Payout_{HM}) - \{(1 - P(W_{HM})) * @Risk_{HM}\}] / @Risk_{HM}} \qquad \mathbf{4.11}$$

This will usually be a constraint rather than an objective. Clearly it is not too wise to have a negative expected value. In many cases the EVROI is negative for both teams. On those occasions the smartest bet one can make is a no bet at all.

[7] Source: http://heritagesports.eu/, accessed December 5, 2014.
[8] http://entertainment.howstuffworks.com/sports-betting2.htm, accessed December 5, 2014.
[9] http://www.sportsbettingonline.net/strategy/implied-probability/, accessed December 5, 2014.

Second is the Edge and is formulated as:

$$\textbf{Investment Edge}_{HM} = P(W_{HM}) - NIP(W_{HM}) \qquad 4.12$$

Again, the Edge is more likely to be a constraint rather than as an objective. A positive Edge is indicative of a favorable investment opportunity. In short, the EDGE is our measure of competitive advantage.

Caution, just because there is a large EVROI or EDGE one must not conclude that it is a good investment. There are situations where bets appear to be too good to be true, and in fact are just that – too good to be true! This and other anomalies will be addressed through our Investment Function.

5 Investment, Staking, Filters, and Decisions

The financial objective is to maximize profits by: sizing @risk capital (limiting investor exposure, etc.)
Subject to: economic and operational constraints

Conceptually one wishes to size investment in relationship to the competitive EDGE and or the Expected Return on Investment. What does that relationship look like?

Joe Peta was an options trader who coupled much of his technical trading expertise with his love of baseball in his book on sports investment [2]. One of his more ingenious observations was the quantifying the level of investment (based on percentage of bankroll) and the magnitude of the betting EDGE. Figure 4 represents the Peta calculated level of investment when the overall probability of winning history is about 55% to 58%. Based upon Expected ROI the "S" curve in Figure 6 is derived and complimentary to that shown in Figure 5 and also is predicated upon the same winning percentage.

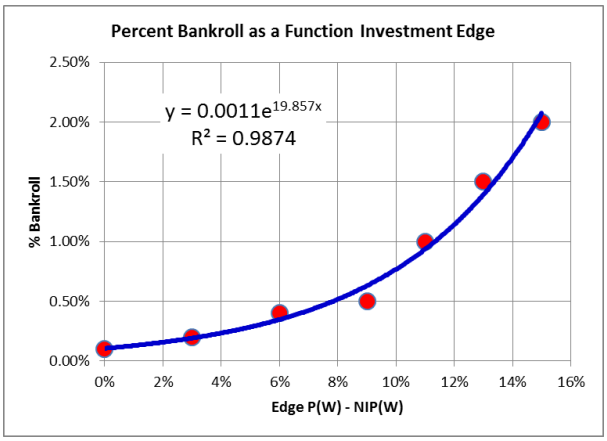


Figure 5 - Investment Tactics Schedule f(EDGE)
Source:[2]

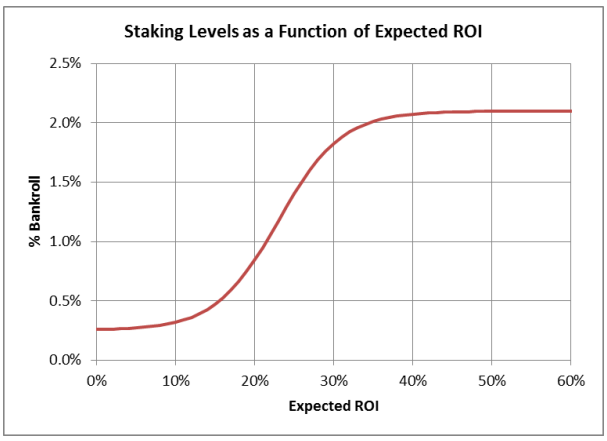


Figure 6 - Staking Strategy f(EVROI)
Source: [3]

The previous curves don't answer the question about whether a bet should be made or not, but only the magnitude if an investment should be undertaken. Since our model has markédly improved the probability of winning, the investment curves (Figures 5 and 6) will be shifted upward. The following is the actual investment function incorporated into our sample game and detailed in Appendix C.

$$\%\text{Bankroll} = \text{Shift} * .0061 * (e^{.484 * EDGE}) \qquad 5.11$$

Subject to: **Maximum investment per bet**
Maximum daily investment

The amount of the shift (7.9 in our model) is derived during the optimization process. A key to improving the probability of winning (see Results section) is picking the low hanging fruit of success by selectively building filters. Unless specific operational, economic and financial criteria are, met an investment will not be allowed. The Edge will be used throughout the study as the primary investment metric since it is literally defines the meaning of Competitive

Advantage. The following table outlines representative boundaries incorporated into the model and applied to raise the probability of winning and the resultant profit levels.

Table 2 – Example of Team Selection Criteria Filters

Criteria	1	2	3	4	5	6	7	8	9	10
	Net Rank	**%Δ EVR (Hm – Rd)**	**%Δ K/BB**	**%Δ OPS**	**NP(W)**	**EDGE Lower**	**EDGE Upper**	**PA_{RD} EDGE**	**PA_{HM} EDGE**	**Dynamic Age DB**
Road	<-10	<-25%	<-50%	<-25%	>47%	>0%	<21%	>71	>11	≈ 65
Home	>9	>15%	>75%	>10%	>45%	>0%	<17%	>10	>52	

The decision to invest is constrained by the filtering boundaries which are recalculated on a daily basis by incorporating a profit maximization (calculated over most recent 30 days) genetic programming algorithm. Each filter has a different value based upon whether it is applied to the road or home team. Let's review each criterion in Table 2:

- 1- Net Rank tabulates the offensive and defensive ranks of the road and home teams from best (1) to (30) worst in MLB. Average runs/game are adjusted, based on league performance. The four ranks are then tabulated in such a way that positive numbers indicate a favorable posture for the home team and negative numbers favor the road team. The ranking calculations will vary with time duration of the database.
- 2- %Δ EVR (Hm – Rd), percent change for Normalized Expected Runs, Home vs. Road, based upon calculations in equation 2.52.
- 3 - %Δ K/BB, calculates the strikeouts/base on balls for each team (road and home) then derives the % difference between the two.
- 4 - %Δ OPS, percent difference between on base and slugging percentages.
- 5 - NP(W), normalized winning percentages, results of Monte Carlo simulation using equation 2.35.
- 6 – EDGE, Lower, minimum acceptable betting edge from equation 4.12.
- 7 - EDGE Upper, maximum level of EDGE, when things look too good to be true as certainly is the case in sports investment. This critical number defines the upper level of an acceptable EDGE, too good to be true.
- 8 - PA_{RD} EDGE, for road team, minimum number of plate appearances in relevant data subsets to be used in calculating matchup statistics. This value is calculated during the optimization process. Without sufficient data no investments will be made.
- 9 - PA_{HM} EDGE, same as 8 above but for home team.
- 10 – Dynamic Age DB, all statistical calculations are predicated upon how far back in time (in days) one goes to back before the data becomes style. This variable is also derived during the optimization process.

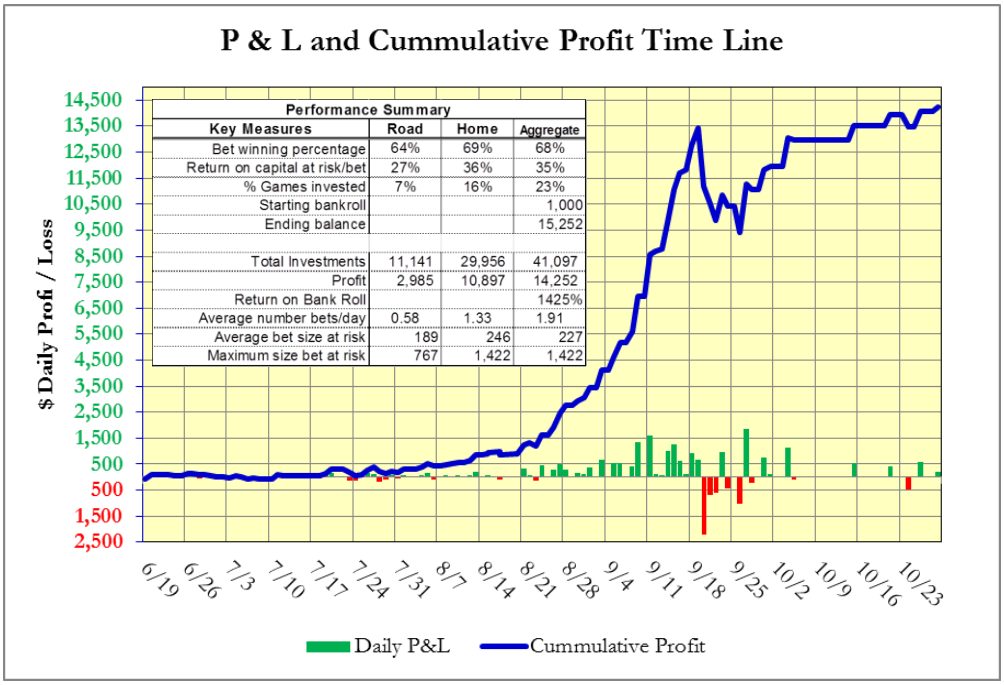
The pieces are now in place, the game, finance, and decisions. It's Time to recap the results.

6 Results

Key concepts have been incorporated into a robust stochastic model including:

- Comparable units Runs/Out linked with in prediction mode (Figure 2)
- Stochastically based probability of winning (equation 2.35),
- Investment selection through criteria filtration (Table 2),
- Staking and investment levels based on risk and return (Figures 5 and 6),
- Nonlinear optimization simultaneously applied to operational and financial variables,
- Integration of micro batter pitcher matchups (equations 2.41-2.44) with macro team performance (Table 2 column 2),
- Time period dependent dynamic data (Table 2 column 10, Appendix C).
- Single equation to assign a productive value to a team or player.
- Normalizing inconsistent data (Table 1 and Equation 2.52).

That's fine, but does it work? Let's let the numbers tell the story[3].



Performance Summary			
Key Measures	Road	Home	Aggregate
Bet winning percentage	64%	69%	68%
Return on capital at risk/bet	27%	36%	35%
% Games invested	7%	16%	23%
Starting bankroll			1,000
Ending balance			15,252
Total Investments	11,141	29,956	41,097
Profit	2,985	10,897	14,252
Return on Bank Roll			1425%
Average number bets/day	0.58	1.33	1.91
Average bet size at risk	189	246	227
Maximum size bet at risk	767	1,422	1,422

Figure 7 – Time Line of Daily and Cumulative Profits, Performance Summary

The Performance Summary in Figure 7 several elements should be noted:

- Only 23% of all games warranted investment.
- Overall winning percent was 68%.
- Average return on capital at risk was 35 %
- In about four months, the bankroll grew by a factor of fourteen.

The methods incorporated have been used in varying degrees for over a decade.

7 Conclusions

Two questions were initially posed and have been addressed.

First, a model was developed to accurately depict the matchup characteristics between the pitcher vs. batter. Concurrently team and park macro influences were brought into play in order to broaden the scope of the prediction model. Using Monte Carlo simulation an accurate estimation of the probability distribution of winning was achieved. Coupling the probability of winning with the betting line yielded the expected economic consequence of the market inequities – the EDGE.

Second, an investment function was extrapolated capturing the elements of risk and resultant financial return. Knowing the EDGE, the investment function provides the mechanism for determining just how much should be put at risk.

Looking forward, a ball club can use the methodologies covered to value batters and pitchers, identify strength and weaknesses his own and competing teams, and provide alternatives to player game day selection and lineups. A Portfolio managers can use the investment function to create and generate a somewhat elusive "efficient frontier".

Bookies, get out of Dodge!

8 Acknowledgements

It is with much personal appreciation that I give thanks to those that have supported and contributed to my understanding of baseball and applying the appropriate mathematics.

From a technical standpoint Dr. Bill Strause (FutureMetrics) has simply been invaluable. My most esteemed protégé Dr. Sola Talabi (Carnegie Mellon University) asks the most probing questions and stimulates innovation. From Palisade Corporation Sam McLafferty and Dr. Javier Ordóñez have always been on call and more than responsive to my most trivial requirements.

Baseball insight is essential. Les Otten (former Vice Chairman, Boston Red Sox) has leaned over backwards to be supportive and provide a unique perspective on the game. Lou Piniella (former manager Chicago Cubs) has shared cryptic insights into and what a manager can and cannot do. Chuck Armstrong (former President Seattle Mariners) allowed for me to meet and discuss the idiosyncrasies of team management of the Mariners.

Jeffery Ma and Mark Kimmel (formerly of Citizens Sports) welcomed me into the San Francisco sports gaming community and provided valuable and useful insights into the game and methods of analysis.

My wife Peggy and daughters Candy and Ashley provide unique personal support and an extraordinary professional perspective on my faddish with baseball.

9 References

[1] Wayne L. Winston, Mathletics,1st Edition, Princeton, New Jersey, 2009.
[2] Joe Peta, Trading Bases, 1st Edition, New York, New York, 2013.
[3] Jeffrey Ma, The House Advantage, New York, New York, 2010.