UCLA College
Master of Applied Statistics

Beating the Book:

A ML Approach to Identifying an Edge in NBA Betting Markets

Guy Dotan

Masters Thesis - 2020

Link to Publication

Updated: March 23, 2022

Thesis Overview

PART 1 – Introduction

PART 2 – The Mathematics of Sports Gambling

PART 3 – Model Building

PART 4 – Beating the Book

PART 5 – Conclusions

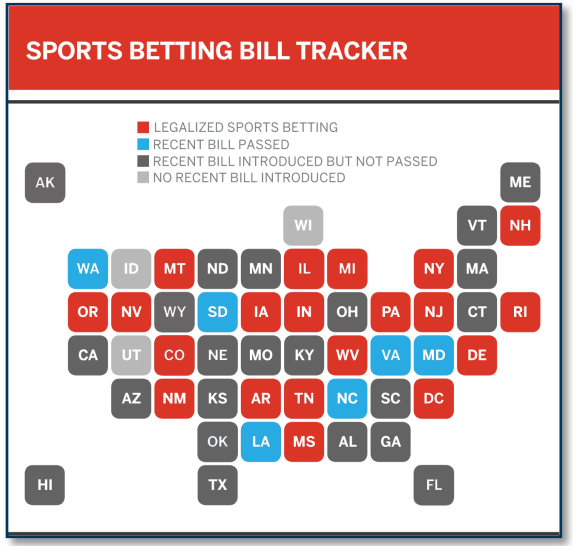
Setting the Stage

PROFESSIONAL SPORTS ANALYTICS

- Quantitative analytics departments within NFL, NBA, NHL, MLB, NHL, MLS organizations have become commonplace
- Sports analytics market expected to reach a revenue of $4.5B by 2024 and growing at a rate of 43.5%.[1]

LEGALIZATION OF SPORTS GAMBLING

- May 2018, Supreme Court overturns *Professional and Amateur Sports Protection Act (PASPA)*
- Each state permitted to make own sports wagering laws
- As of May 2020, 24 states have passed or pending gambling legislation and sportsbooks have generated $1.4B revenue[2]



Updated as of Dec. 2020 [3]

Closing the Gap for Bettors

Casinos have had the upper hand, but analytics is leveling the playing field

Proliferation in gambling

Legalization in more states has led to an increased interest in gambling to all sports fans from the casual to the passionate.

Rise in Analytics

The average fan is far more more analytically inclined than before and consequently so is the average bettor.

Data Democracy

Open-source communities with stat archives, professional leagues sharing their own data through APIs, and the rise of sports analytics conferences and blogs has lowered the barrier to enter.

Sports Betting 101

NBA Game Lines			
Team	Money	Spread	Total
January 18, 2020			
GOLDEN STATE WARRIORS	-235	-5 (-110)	O 227 (-110)
LOS ANGELES CLIPPERS	+185	+5 (-110)	U 227 (-110)

TOTALS

The total (over/under) is the expected combined score of both teams, regardless of who wins. Bettor must predict if the total game score is above or below that betting line.

MONEYLINES

Bettor must predict the winner of the game, regardless of margin of victory.

SPREADS

The spread is the projected margin of victory between the two teams. Bettor predicts the winner of the game after including the spread to the final score.

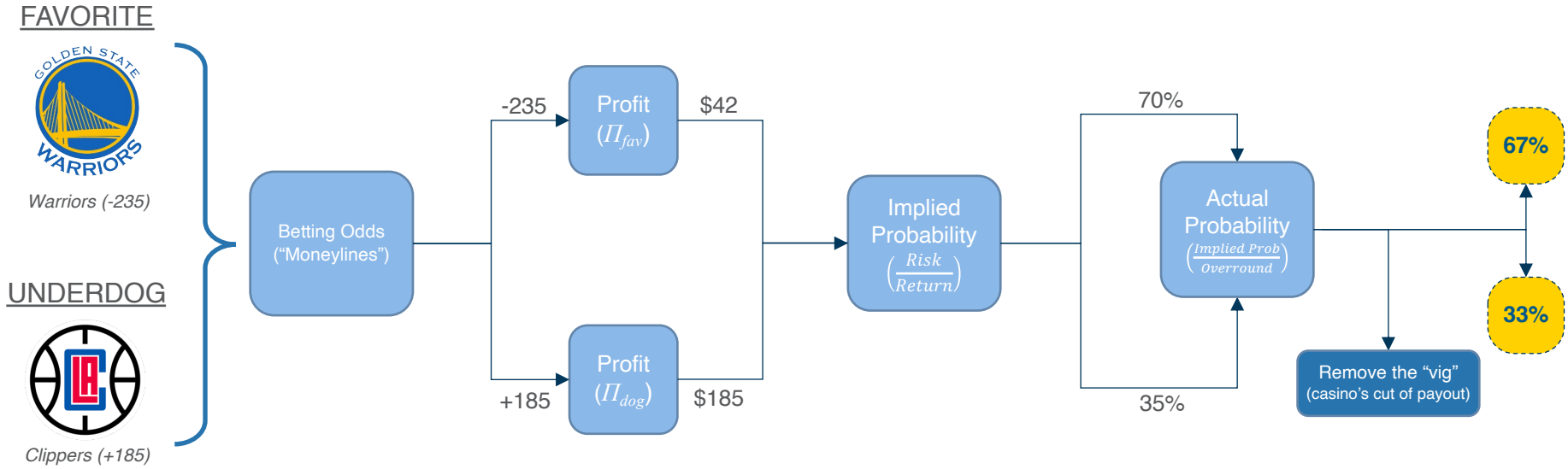
ODDS

The payout of winning the bet is indicated by a negative number (-235) for the favorite and a positive number (+185) for the underdog. The profit formulas for each payout are shown in the notes on the next slide: Π_{fav} , Π_{dog}

The Mathematics of Sports Gambling

DERIVATION FROM BETTING ODDS TO ACTUAL WIN PROBABILITY[4]

(using example of risking $100 on a moneyline wager)



Notes

$$\Pi_{fav} = \frac{100}{-1 * ML} * Risk$$

$$\Pi_{dog} = \frac{ML}{100} * Risk$$

$$Return = Risk + Profit$$

$$Overround = IP_{fav} + IP_{dog}$$

Methodology Overview

BUILD MODEL

Create model to predict each team's chances of winning a matchup on a 0-100% scale

CONVERT BETTING LINES

For a sample of NBA games, convert each team's betting lines to actual win probabilities

PLACE BETS

Use model output each game to determine if there is an edge to bet on and how much to wager

Data Sets

GAMBLING DATA

- Betting lines archive for every NBA game from 2007-08 to 2019-20

NBA TEAM REGULAR SEASON GAME LOGS

- `nbastatR` package in R with vast API of NBA statistics

- Available team data:
  - Basic statistics (points, rebounds, assists, etc.)
  - Matchup data (home/away, days of rest, game number of the season)
  - Advanced metrics (efficiency ratings, pace of play)

EDA & Data Processing

- Year-over-year data suggests a positive trend in pace (possessions per game)
- Scaled box score stats to per-100 possessions to standardize the data and make comparable between seasons

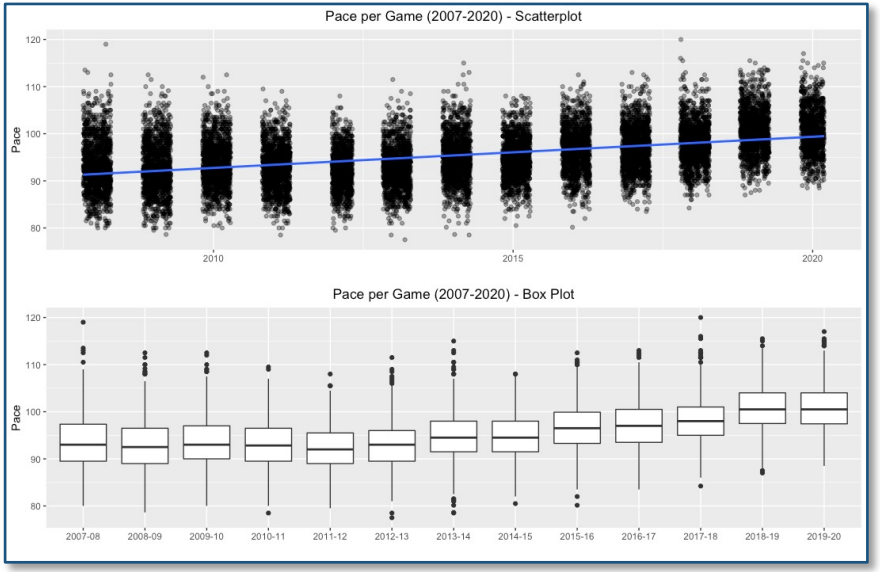


Figure 1 – Eminent positive trend in pace per game over the past 13 seasons.

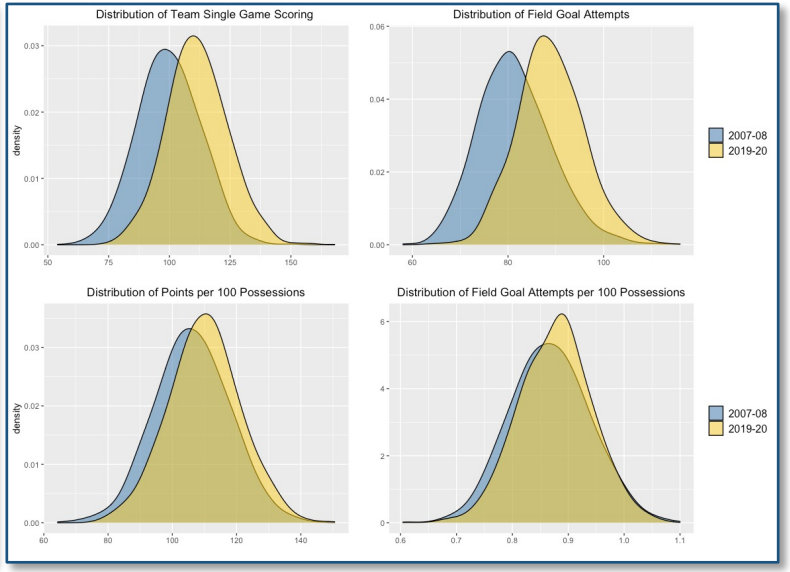
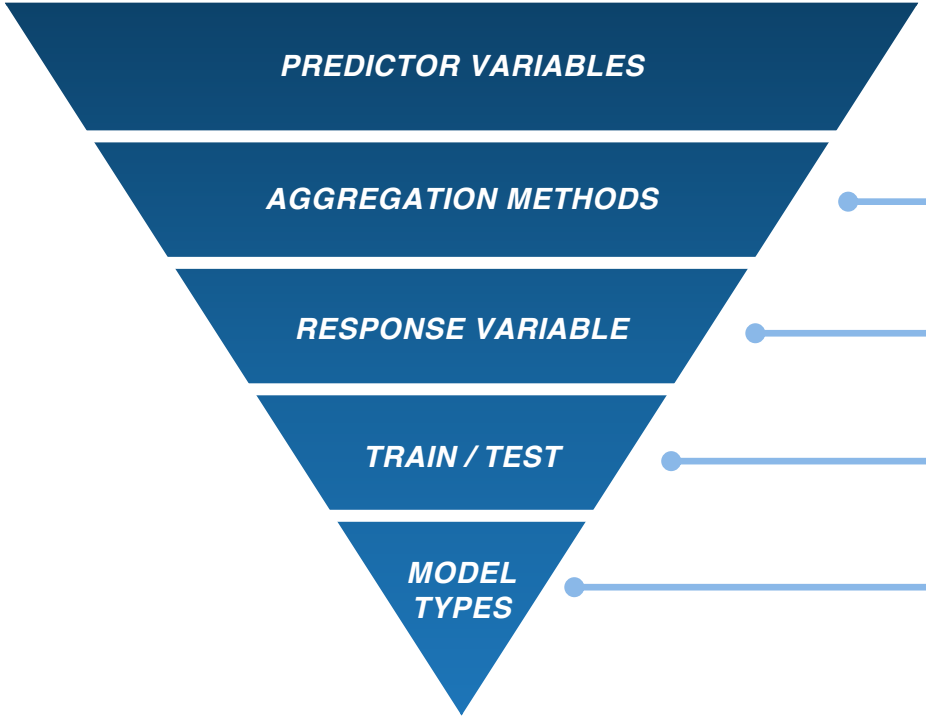


Figure 2 – Distributions of team points and FGA for the 2007-08 vs. 2019-20 seasons when scaled per game and per 100 possessions.

Model Structure



- Own box score per-100 possessions metrics
- Opponent box score per-100 possessions metrics
- Matchup variables (home/away, days rest, etc.)

- Applied only to box score metrics

1. All games of season until prediction date*
2. Rolling aggregation of stats over last 8 games*

- Binary response variable (Win or Loss) for each team in each matchup

- 75% / 25% - train/test split
- 10,830 regular season games from 2007 to 2019
- Standard 10-fold, cross-validation

1. Logistic
2. Random Forest
3. XGBoost
4. Neural Networks

** Logic behind aggregation methods and handling of edge cases are addressed in thesis Chapter 4.3*

Assessing Model Accuracy

CONFUSION MATRIX

- Model outputs a confidence score (0-100%) that a team will win
  - Confidence > 50% → Coded as "Predicted - Win"
  - Confidence < 50% → Coded as "Predicted - Loss"
- Predictions from each model grouped into a confusion matrix to determine accuracy

	Actual - Win	Actual - Loss
Predicted - Win	***True Positive***	***False Positive***
Predicted - Loss	***False Negative***	***True Negative***

PERFORMANCE METRICS

- Primary metrics used when assessing model accuracy:
  - **Accuracy** – the ratio of correct predictions to total possible
  - **Kappa** – observed accuracy compared to expected accuracy
- F1 is a common metric used for binary response variables, calculated from precision and recall:
  - **Precision** – % of positive predictions that were truly positive
  - **Recall** – % of the true positives predicted as positive

Primary

$$Accuracy = \frac{TruePos + TrueNeg}{Total}$$

$$Kappa = \frac{ObsAccuracy - ExpAccuracy}{1 - ExpAccuracy}$$

Additional

$$Precision = \frac{TruePos}{TruePos + FalsePos}$$

$$Recall = \frac{TruePos}{TruePos + FalseNeg}$$

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

Model Summary

MODEL COMPARISON

- Two feature selection methods used for the logistic model
  - Recursive Feature Elimination (RFE) & Stepwise Selection (Step)
- All models yielded at least **60% accuracy**, thus predicted better than chance
- Year-to-date aggregation method outperformed 8-game span in each model type

Model Type	Aggregation Type	Accuracy	Accuracy (95% CI)	Kappa
Logistic - RFE	8-game span	0.647	[0.6316, 0.6631]	0.295
Logistic - RFE	Year-to-date	0.659	[0.6436, 0.6747]	0.319
Logistic - Step	8-game span	0.616	[0.5994, 0.6315]	0.231
Logistic - Step	Year-to-date	0.622	[0.6055, 0.6373]	0.243
Random Forest	8-game span	0.637	[0.6207, 0.6524]	0.273
Random Forest	Year-to-date	0.649	[0.6328, 0.6641]	0.297
XGBoost	8-game span	0.638	[0.6218, 0.6535]	0.275
XGBoost	Year-to-date	0.659	[0.6433, 0.6745]	0.318

BEST MODEL

✓ **Logistic (RFE) model** using year-to-date stats had the best accuracy and best F1. Performed slightly better than the complex machine learning techniques: Random Forest, XGBoost, and Neural Networks*

Model Type	Aggregation Type	Precision	Recall	F1
Logistic - RFE	8-game span	0.6480	0.6510	0.6490
Logistic - RFE	Year-to-date	0.6570	0.6670	0.6620
Logistic - Step	8-game span	0.6140	0.6286	0.6212
Logistic - Step	Year-to-date	0.6203	0.6271	0.6237
Random Forest	8-game span	0.6353	0.6469	0.6410
Random Forest	Year-to-date	0.6490	0.6460	0.6480
XGBoost	8-game span	0.6373	0.6447	0.6409
XGBoost	Year-to-date	0.6600	0.6570	0.6580

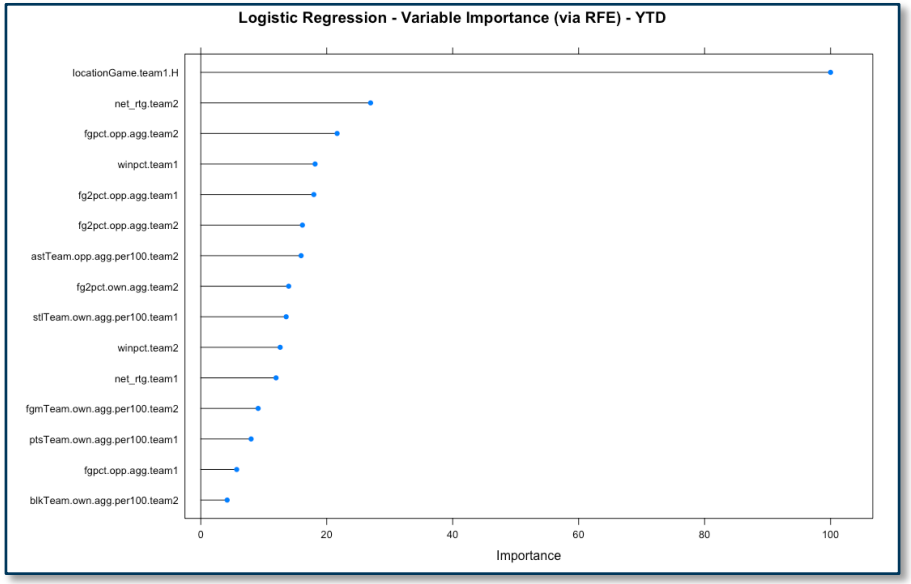
** Neural network results not shown in table but discussed in thesis Chapter 6*

Most Performant Model – Logistic Regression

65.9% *ACCURACY*

31.9% *KAPPA*

66.2% *F1*



VARIABLE IMPORTANCE

- High leverage variables in logistic model (and most other models)
  - Location: Home vs. Away
  - Net Rating: Offensive Rating - Defensive Rating
  - Team Winning Percentage
  - Field Goal Percentage
  - Assists per-100 Possessions

CONCLUSION

- Although most models performed similarly, the logistic model is far more interpretable than the black-box ML methods
- The **logistic model with year-to-date aggregated statistics** was implemented in the final section of the thesis in which we applied a model using statistics from the 2019-20 NBA season to the betting odds for those games

Beating the Book

SPORTSBOOK OBJECTIVES

- The casino takes a cut ("vig") of all bets in the book. Thus, to ensure profitability on a line, they need an **even amount of money** wagered on both sides of the bet.
- Generally, that means bookmakers set a line that accurately projects true outcome of the game
- But bookmakers must also account for inherent biases from bettors
  - More volume of bets from the hometown teams of **larger markets**
  - Wagering on more exciting outcomes (**high scoring** games, **underdog** upsets)
  - Favoring the teams of well known or **superstar** players

FINDING AN EDGE

- With bettor biases influencing sportsbook lines, the access to data, and a powerful machine leaning model, **can we build a model that identifies an edge in a betting matchup?**
- Additionally, if we applied this model to an optimized betting strategy can we **systematically beat the market** over the course of a season of betting?

Model Application

Applied our model to each game of 2019-20 NBA season to get a projected win probability for each team in the matchup. Then used various betting strategies to determine whether and how much to wager on the game.

FIXED BET

- Compare model probability to betting line's true probability
- Wager same amount ($10) on each game's moneyline
- End of the season see if final bankroll is positive or negative

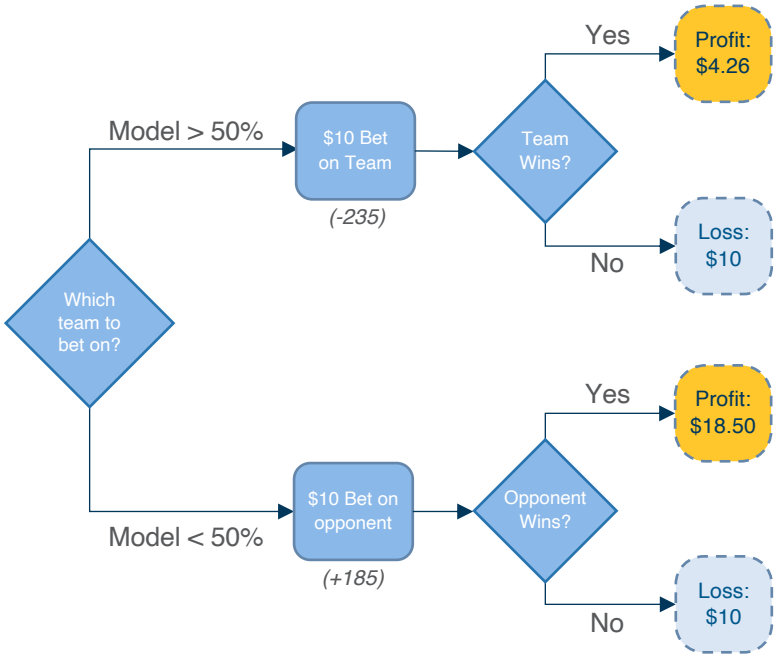
KELLY CRITERION

- Compare model probability to betting line's true probability
- Use Kelly Criterion to determine how much to wager on each moneyline bet
- See how $10,000 initial bankroll fluctuates over course of the season

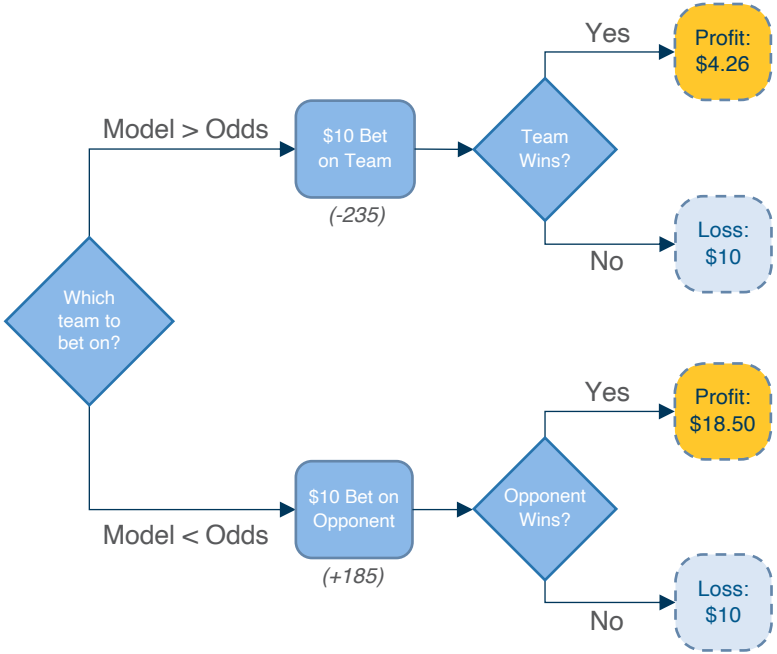
Model-Based Fixed Bet Logic Flows

Core fixed wagering strategies stem from comparing our model's predicted odds to betting line odds

FIXED BET – MODEL FAVORITE



FIXED BET – MODEL EDGE



Bankroll over Time using Fixed Bet Implementations

Simulated the model's success by applying six wagering strategies on each game of the 2019-20 season

MODEL EDGE
Only bet if model odds > betting odds

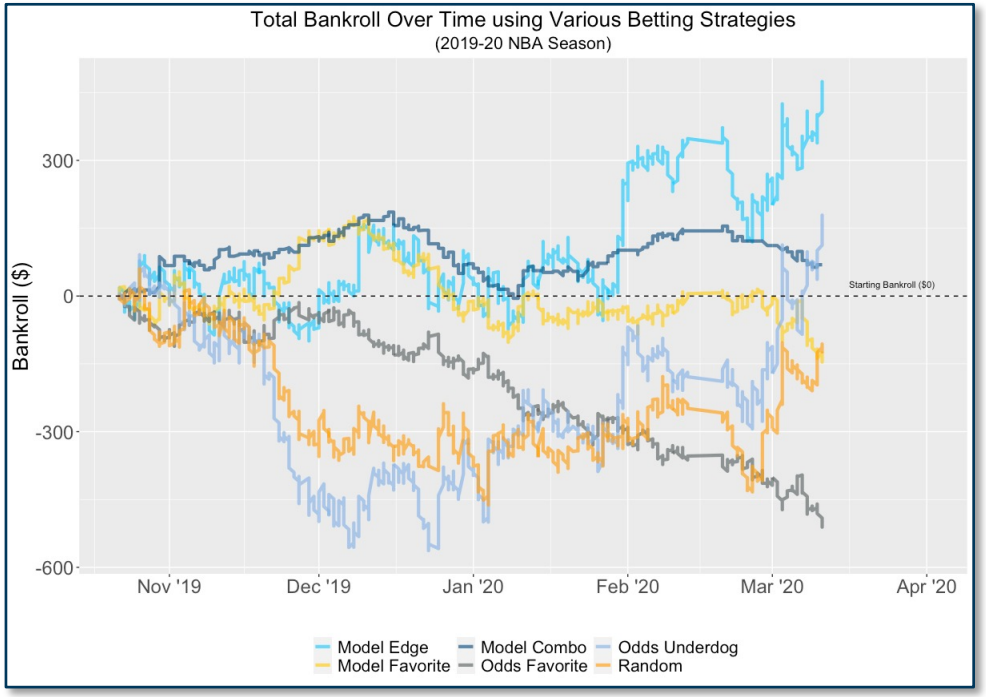
MODEL FAVORITE
Only bet if model predicts team will win

MODEL COMBO
Only bet if model odds > betting odds ***and*** *model predicts team will win*

ODDS FAVORITE
Bet on the team more likely to win (model agnostic)

ODDS UNDERDOG
Bet on the team less likely to win (model agnostic)

RANDOM
Random team selection in each matchup; used as a control (model agnostic)



Fixed Betting Results

OVERVIEW

- Each betting strategy involved wagering up to $10 on each game
  - Not betting on the matchup occurred if model did not determine an edge versus the betting line
- 971 total games in test dataset, so maximum bankroll loss of $9,781

SUMMARY

- **Model Edge** was most successful strategy, yielding final bankroll of **$464** (**+4.8% ROI**), despite accurately picking winner just **44%** of the time
- **Model Favorite** picked correct winner of the matchup **66%** of the time. But ended down **$145** by end of season (**-1.5% ROI**)
- **Model Combo** was very conservative on when to wager, betting on just **23%** of all possible matchups. But still ended with a positive ROI (+**3.1%**).
- Model agnostic strategies lost money when picking favorites (-5.3%) or randomly (-1.2%). But yielded a positive ROI (+1.8) when selecting only underdogs.

Table 7.1: Summary results between fixed betting methodologies.

	Win	Loss	No Bet	Win Pct	Final Bankroll	ROI	Bankroll SD*
Model Edge	430	541	-	44.3%	$464.17	4.8%	438.72
Model Favorite	645	326	-	66.4%	$-144.77	-1.5%	244.15
Model Combo	142	84	745	62.8%	$69.23	3.1%	129.20
Odds Favorite	645	309	17	67.6%	$-509.81	-5.3%	213.68
Odds Underdog	309	645	17	32.4%	$168.79	1.8%	507.96
Random	474	497	0	48.8%	$-116.14	-1.2%	292.30

*Final bankroll standard deviation calculated by running a simulation 1,000 times using a randomized game outcome based on the betting odds win probability as the "true probability" for each game.

The Kelly Criterion

CORE MOTIVIATION

- In practice, bettor should not wager same amount on each game
- Some matchups carry higher degree of confidence than others, i.e., exists a **larger perceived edge**
- Methodology to determine how much to wager in each game: **The Kelly Criterion**
  - Definition: "the mathematical formula for investors and gamblers that calculates what percentage of their budget they should allocate to each investment or bet."[5]
  - Informs a bettor how much more to wager on high-confidence games and the extent to avoid bets with low confidence, thus resulting in **limiting losses** and **maximizing gains**.

EQUATION

(Derivation of Kelly Criterion formula included in thesis Chapter 8.1)

$$k = \frac{pb - q}{b}$$

k: the Kelly Criterion – optimal proportion of bankroll to invest
p: the true probability of an event
q: probability the event does not occur ($q = 1 - p$)
b: given odds of the event occurring (in the format of *"b to 1"*)

A Kelly Criterion Example

SAMPLE SCENARIO

vs.

- Warriors favored versus Clippers (-110)
- Model projects Warriors have a 55% chance to win
  - Therefore, Clippers have 45% chance to win
- Moneyline → Odds conversion: 100/110 = .909

Moneyline to Odds Conversion - General Equation

$$\text{Odds} = \begin{cases} \dfrac{ML}{100}, & \text{if ML} \geq 0 \\ \dfrac{100}{-ML}, & \text{if ML} < 0 \end{cases}$$

SOLVING FOR K

p: true probability of the event = **55%**

q: probability the event does not occur = **45%**

b: odds of given event occurring = **.909**

$$k = \frac{pb - q}{b}$$

$$k = \frac{0.55 \times 0.909 - 0.45}{0.909}$$

$$k = \frac{0.04995}{0.909}$$

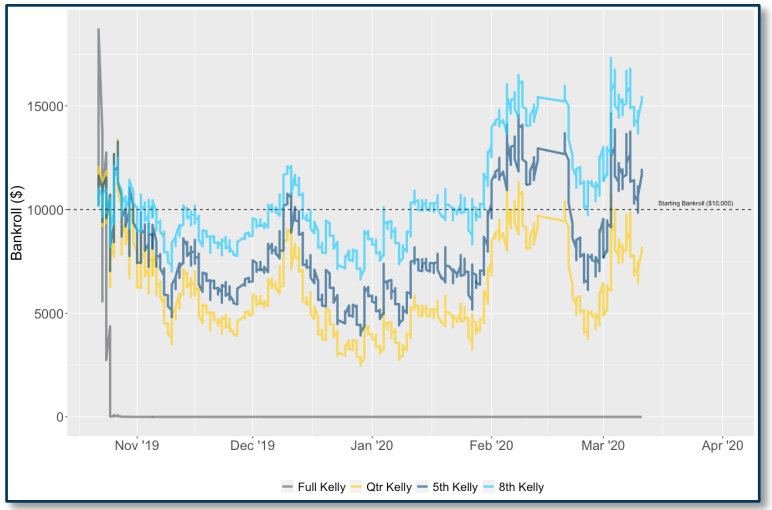
$$k = 0.05495$$

Optimal wager amount is **5.5%** of total bankroll

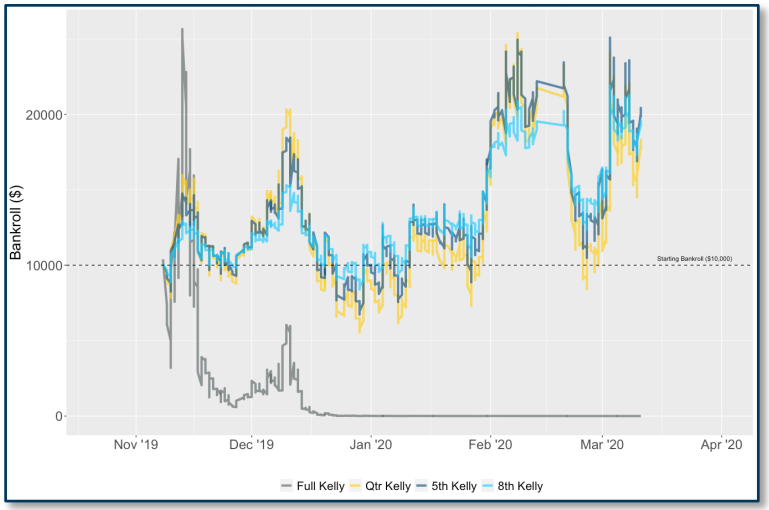
Results from Kelly Implementation

Applying same prediction model over the 2019-20 NBA season but with various fractional Kelly Criteria

- Start with $10,000 bankroll, use "model edge" methodology but with **variable wager amounts** based on Kelly Criterion
- Common critique of Kelly strategy is **over-aggressiveness** that leads to wild bankroll **fluctuations**
  - One of the recommendations to smooth out the wild oscillations common to Kelly-based wagering is to use the concept of "**fractional Kellys**"[6]. E.g., instead of a full Kelly wager, use half-Kelly (*k/2*) or quarter-Kelly (*k/4*)



BANKROLL OVER TIME USING VARIOUS KELLY CRITERIA



BANKROLL AFTER EXCLUDING EACH TEAM'S FIRST EIGHT GAMES OF THE SEASON

The Lucrative Potential of Fractional-Kelly Wagering

CONCLUSIONS AFTER FULL SEASON

- **Full Kelly** strategy proved far too aggressive, drove bankroll to $0 in first month of betting
- **Quarter Kelly** still too aggressive, losing about **$2,100** from initial bankroll (**-21% ROI**).
- Best results came from **8th Kelly**, yielding gain of about **$5,000** (**+52% ROI**), despite the same **44.3%** accuracy as the fixed bet - model edge method.

CONCLUSIONS AFTER EXCLUDING EARLY SEASON

- Insufficient data has accumulated until each team's eighth game is played in order to meet the specifications of our year-to-date model.
- **Full Kelly** still rapidly trends to $0.
- **4th, 5th,** and **8th Kelly** methods all yielded high ROIs. The best being 5th Kelly which **nearly doubled** the initial $10K bankroll (**+98% ROI**).

Using the same predictions as the fixed bet - model edge method, the Kelly Criteria technique increased our maximum ROI from 4.8% with fixed betting to 98% with 5th Kelly!

Table 8.1: Summary results between our fractional Kelly criteria.

	Win	Loss	No Bet	Win Pct	Final Bankroll	ROI	Bankroll SD*
Full Kelly	430	541	-	44.3%	$0.00	-100%	<0.01
Qtr Kelly	430	541	-	44.3%	$7,865.77	-21.3%	5.557.94
5th Kelly	430	541	-	44.3%	$11,565.00	15.6%	5,662.36
8th Kelly	430	541	-	44.3%	$15,150.70	51.5%	4,897.09

*Final bankroll standard deviation calculated by running a simulation 1,000 times using a randomized game outcome based on the betting odds win probability as the "true probability" for each game.

Table 8.2: Fractional Kelly results after excluding each team's first eight games.

	Win	Loss	No Bet	Win Pct	Final Bankroll	ROI	Bankroll SD*
Full Kelly	375	471	-	44.3%	$0.02	-100%	2.96
Qtr Kelly	375	471	-	44.3%	$17,369.64	76.4%	7,067.40
5th Kelly	375	471	-	44.3%	$19,824.89	98.2%	6,301.83
8th Kelly	375	471	-	44.3%	$19,186.95	91.9%	4,755.52

*Final bankroll standard deviation calculated by running a simulation 1,000 times using a randomized game outcome based on the betting odds win probability as the "true probability" for each game.

Conclusion

FURTHER RESEARCH

- Explore more advanced stats beyond per-possession metrics. E.g., tracking metrics.
- Incorporate strength of schedule into the model
- Include player performance (and injuries) into the model, to supplement team stats
- Building a model to predict spreads instead of moneylines would allow for increased likelihood to win when betting on underdogs.

FINAL THOUGHTS

- A simpler **logistic** model performed as well on training set as the ML-intensive techniques
- Model achieved **65% accuracy** predicting results on a full season of NBA games it had never seen
- Identifying an “edge” then applying fractional-Kelly Criteria achieved nearly **2x ROI** over the season
- Our study demonstrates that with sufficient data, a well tuned model, and optimized betting strategies **it is quite possible to “beat the book”!**

Appendix

BEATING THE BOOK: A ML APPROACH TO IDENTIFYING AN EDGE IN NBA BETTING MARKETS

References

[1] Infoholic Research LLP. “Sports Analytics Market Forecasts up to 2024.” Technical report, November 2018. *[link]*

[2] LegalSportsBetting.com. “Legal Sports Betting Revenue Tracker.” March 2020. *[link]*

[3] Ryan Rodenberg. “United States of sports betting: An updated map of where every state stands.” *ESPN*, Dec. 2020. *[link]*

[4] Sports Betting Dime. “Beating the Juice: Removing the Vig from Sports Betting Lines.” *Sports Betting Dime*, February 2020. *[link]*

[5] Justin Kuepper. “Using the Kelly Criterion for Asset Allocation and Money Management.” *Investopedia*, March 2020. *[link]*

[6] Edward O. Thorp. “The Kelly Criterion in Blackjack Sports Betting and the Stock Market.” *Finding the Edge: Mathematical Analysis of Casino Games*, 1997. *[link]*

Data Dictionary

Variables used in our modeling process, along with their categories and descriptions

Category	Variable	Description
Matchup	idGame	(num) Unique game ID
	slugSeason	(num) NBA season
	gameDate	(date) Date of the game
	beforeASB	(T/F) Game is before All-Star Break (T) or after (F)
	locationGame	(char) Home or away
	slugMatchup	(char) Game matchup
	slugTeam	(char) Team abbreviation
	slugOpponent	(char) Opponent team abbreviation
	numberGameTeamSeason	(num) Team's Xth game of season
	isB2B	(T/F) Is the team on a back-to-back (i.e. 0 days rest)
	isB2BFirst	(T/F) First game of a back-to-back
	isB2BSecond	(T/F) Second game of a back-to-back
	countDaysRestTeam	(num) Number of days since the team's last game
	countDaysNextGameTeam	(num) Number of days until the team's next game

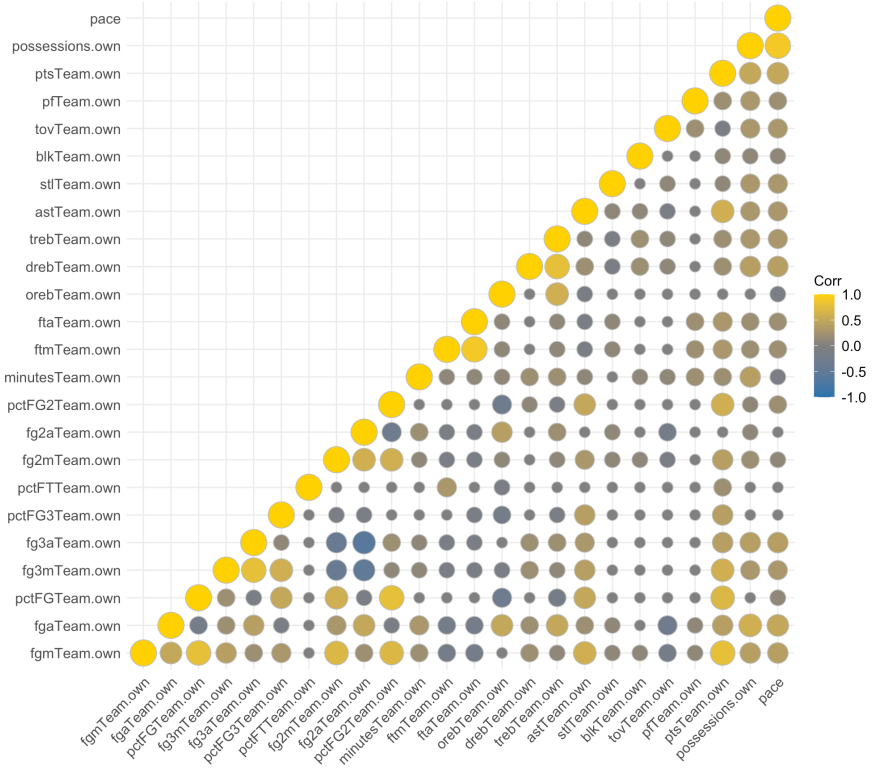
Data Dictionary (cont.)

Category	Variable	Description
Outcome	slugTeamWinner	(char) Winning team abbreviation
	slugTeamLoser	(char) Losing team abbreviation
	outcomeGame	(char) Team win or loss
	isWin	(T/F) Team win or loss
Betting Odds	teamML	(num) Moneyline odds for the team
	oppML	(num) Moneyline odds for the opponent
	teamWinProb	(num) Win probability for the team
	oppWinProb	(num) Win probability for the opponent
Traditional Stats	fgmTeam	(num) Field goals made by the team
	fgaTeam	(num) Field goals attempted by the team
	pctFGTeam	(num) Field goal percentage from team
	fg3mTeam	(num) 3PT field goals made by the team
	fg3aTeam	(num) 3PT field goals attempted by the team
	pctFG3Team	(num) 3PT field goal percentage by the team
	pctFTTeam	(num) Free throw percentage by the team

Data Dictionary (cont.)

Category	Variable	Description
Traditional Stats	fg2mTeam	(num) 2PT field goals made by the team
	fg2aTeam	(num) 2PT field goals attempted by the team
	pctFG2Team	(num) 2PT field goal percentage by the team
	minutesTeam	(num) Minutes played by the team
	ftmTeam	(num) Free throws made by the team
	ftaTeam	(num) Free throws attempted by the team
	orebTeam	(num) Offensive rebounds by the team
	drebTeam	(num) Defensive rebounds by the team
	trebTeam	(num) Total rebounds by the team
	astTeam	(num) Total assists by the team
	stlTeam	(num) Total steals by the team
	blkTeam	(num) Total blocks by the team
	tovTeam	(num) Total turnovers by the team
	pfTeam	(num) Personal fouls committed by the team
	ptsTeam	(num) Total points scored by the team
	plusminusTeam	(num) Team score differential (Own - Opponent points)
Advanced Stats	possessions	(num) True team possessions in a game
	pace	(num) Number of possessions per 48 minutes by a team.

Variable Correlation Matrix



Betting Line Prediction Accuracy (2007-08 – 2019-20)

Table 10.1: Record for matchup favorites based on the sportsbook betting odds.

Season	Games	Win	Loss	Even Odds	Win Pct
2007-08	1229	846	363	20	0.700
2008-09	1230	859	358	13	0.706
2009-10	1230	863	365	2	0.703
2010-11	1230	842	383	5	0.687
2011-12	990	677	311	2	0.685
2012-13	1229	841	380	8	0.689
2013-14	1230	841	378	11	0.690
2014-15	1230	863	364	3	0.703
2015-16	1230	851	367	12	0.699
2016-17	1230	810	403	17	0.668
2017-18	1230	837	383	10	0.686
2018-19	1230	818	394	18	0.675
2019-20	971	645	309	17	0.676

Final logistic model from the year-to-date dataset

```
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.8167  -1.0183  -0.1463   1.0149   2.7019

Coefficients:
                            Estimate Std. Error z value             Pr(>|z|)
(Intercept)                -0.664874   1.159278  -0.574             0.566290
locationGame.team1.H       -0.896211   0.042425 -21.125 < 0.0000000000000002 ***
net_rtg.team2               0.077939   0.013344   5.841        0.00000000519 ***
net_rtg.team1              -0.037474   0.013882  -2.700             0.006944 **
winpct.team1               -1.170826   0.292961  -3.997        0.00006428004 ***
winpct.team2                0.807796   0.284718   2.837             0.004551 **
ptsTeam.own.agg.per100.team1 -0.025395   0.013572  -1.871             0.061314 .
ptsTeam.own.agg.per100.team2  0.006667   0.012497   0.534             0.593668
stlTeam.own.agg.per100.team1 -0.066672   0.021957  -3.037             0.002393 **
blkTeam.own.agg.per100.team2  0.026313   0.024454   1.076             0.281905
fgpct.own.agg.team2        -2.127449   2.769143  -0.768             0.442327
fgpct.own.agg.team1         2.451518   2.760752   0.888             0.374547
fg2pct.own.agg.team2        5.463032   1.750985   3.120             0.001809 **
fg2pct.own.agg.team1       -1.790129   1.751866  -1.022             0.306856
fg2pct.opp.agg.team2       -7.814259   2.185005  -3.576             0.000348 ***
fgmTeam.own.agg.per100.team1 -0.004568   0.022912  -0.199             0.841967
astTeam.opp.agg.per100.team2 -0.051764   0.014653  -3.533             0.000411 ***
fgmTeam.own.agg.per100.team2 -0.048245   0.022915  -2.105             0.035253 *
fgpct.opp.agg.team1        -4.496422   3.229357  -1.392             0.163814
fg2pct.opp.agg.team1        8.613108   2.177498   3.956        0.00007637280 ***
fgpct.opp.agg.team2        14.625446   3.092135   4.730        0.00000224647 ***
---
Signif. codes:  0 `***' 0.001 `**' 0.01 `*' 0.05 `.' 0.1 ` ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 15076  on 10874  degrees of freedom
Residual deviance: 13195  on 10854  degrees of freedom
AIC: 13237
```

UCLA

THANK YOU

© Guy Dotan - March 23, 2022