UNIVERSITY OF CALIFORNIA

Los Angeles

Beating the Book: A Machine Learning Approach to NBA Win Probabilities in Search of an Edge Over the Betting Odds

A thesis submitted in partial satisfaction of the requirements for the degree Master of Applied Statistics

by

Guy Dotan

2020

© Copyright by Guy Dotan 2020

ABSTRACT OF THE THESIS

Beating the Book: A Machine Learning Approach to NBA Win Probabilities in Search of an Edge Over the Betting Odds

by

Guy Dotan  
Master of Applied Statistics in

University of California, Los Angeles, 2020

Professor Frederic R. Paik Schoenberg, Chair

In this study, we set out to better understand the voter behavior of different demographic partitions of Los Angeles County (LAC). We developed a methodology to enrich and analyze LAC voter data by joining it with US census data and modelling vote by mail rates and turnout rates using general additive models. In doing so, we were able to forecast where in LAC we expect to see the highest rates of turnout by vote by mail and turnout in person for the 2020 presidential election. These findings are timely: In 2020, LAC will roll out a new and novel voting system called Voting Solutions for All People. The findings of the current study provide guidance on how to best allocate resources to high-turnout populations and target outreach to low-turnout populations in preparation for the first election under the new system. This paper presents a combination of visualizations, predictive models, and summary statistics that can inform LAC of where to focus outreach to expand the electorate and build a more representative electorate while also improving preparation in areas where turnout has been historically high.

Chargeback fraud is a massive problem for e-commerce businesses. Using historical ticket order data, several machine learning models are trained and tested to predict which transactions are high risk for chargeback. The results of this thesis show that many fraudulent transactions can be successfully identified and stopped before they are processed. Using these types of models could significantly reduce chargebacks, saving companies time and money.

The thesis of Guy Dotan is approved.

Vivian Lew

Zili Liu

Frederic R. Paik Schoenberg, Committee Chair

University of California, Los Angeles

2020

Table of Contents

1. Introduction
2. The Mathematics of Sports Gambling
3. Exploratory Analysis
4. Random Forest
5. Neural Network
6. Betting Results
7. Kelly Criterion
8. Conclusions
9. References

List of Figures

List of Tables

Acknowledgements

Chapter 1

Introduction

Over the past decade there have been two seminal shifts in the world of sports analytics and thusly the entire professional sports landscape. The first: the proliferation and democratization of accessible data. The second, and more recently: the federal legalization of sports gambling within the United States.

While the sports industry might be one of the newest sectors to be disrupted by the emergence of data-driven decisions challenging pre-conceived notions from “experts”, its impact has been fast and far reaching. Setting aside the unprecedented shock to the economic ecosystem—specifically within sports and entertainment—from the 2020 outbreak of COVID-19, the sports analytics business has been thriving. “The global sports analytics market is expected to reach at a revenue of $4.5 billion by 2024, growing at a CAGR [Compound Annual Growth Rate] of 43.5%” [1].

1.1 – Current State of Sports Analytics

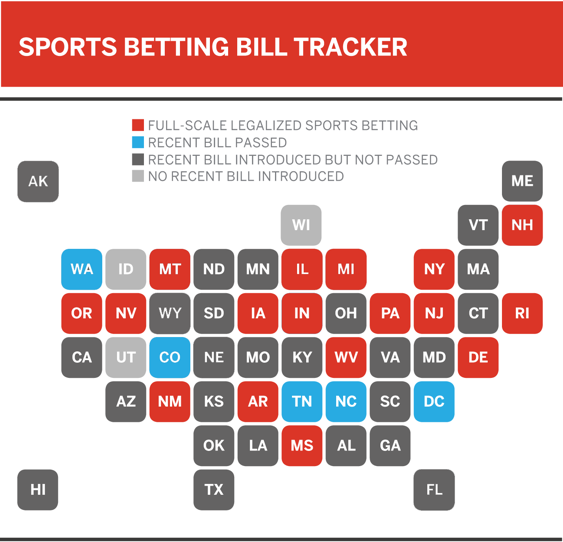
Obviously, the most well-known adoption of analytics into the sports universe was within Major League Baseball, thanks largely to Michael Lewis’ 2003 book *Moneyball* and subsequent movie blockbuster, staring Brad Pitt, in 2011. The book which chronicles the influence of Bill James, the field of empirical baseball research known as Sabermetrics, and the story of the 2002 Oakland A’s success has been the poster child of how data can create a competitive edge on the playing field. But sports analytics has made its impression in far more avenues than just baseball. The field is responsible for the increased emphasis of the three-point shot in basketball, the use of optical player tracking technology in the NFL, and even the statistical optimization of curling game strategy that led the Swedish women’s national team to a gold medal in the 2018 Winter Olympics [2], just to name a few.

The integration of data analysts and scientists as a crucial element of professional sports organizations appears here to stay, but the acceleration in the field’s adoption can be tied to the increased availability of data. The NFL and NBA hold yearly hackathons to allow anyone the opportunity to dive into their sport’s data and present findings to top league officials with prizes and networking at stake. Conferences such as Sloan Sports Analytics Conference in Boston began as a small gathering of about 100 attendees in 2006, and now in 2020, attracts over 4,000 people. The conference has gained national recognition, notably hosting former President Barack Obama as the keynote speaker in 2018. The industry’s explosion in popularity, though, has been aided by communities such as FiveThirtyEight, Retrosheet, Sports-Reference, and league-offered APIs bestowing data democracy to anyone that desires. Sports analytics has largely become open-source and this hive mind has benefited players, teams, and organizations consequently.

1.2 – The Legalization of Sports Gambling

On May 14, 2018, the Supreme Court case *Murphy v. National Collegiate Athletic Association* reached a landmark decision regarding the federal government’s right to control a state’s ability to sponsor sports betting. In a 6-3 decision, the Professional and Amateur Sports Protection Act of 1992 (PASPA) was overturned opening the doors for every state to make its own laws permitting in-state sports wagering.

In just two years since the ruling there are already 17 states with full-scale legalization and another five that have passed legislation that will take effect in the coming year. [3] And as one would expect, bettors in legal states have flocked to sportsbooks, both digital and brick-and-mortar. Since the overturning of PASPA, Americans have placed over $20 billion of bets which has generated $1.4 billion of revenue in those legal states. [4] Morgan Stanley projects that in just five years, by 2025, almost three-quarters of US states (36) will have legalized sports betting and the U.S. market could see $7 to $8 billion in revenue. [5]



1.3 – The Intersection of Data and Wagering

Sportsbook operators within casinos have had decades of experience building a complex infrastructure of analytics to help them determine where to set up their gambling lines. Their goal is to set up a bet every game such that there is an even amount of money wagered on both sides of the bet. This allows them to take their cut of the wagers (known in the industry as the “vigorish” or “vig”) and thus drive revenue to their casino, no matter which team wins. For the entirety of their existence, sportsbooks have maintained a significant edge over the majority of bettors because of their access to data and domain expertise building models to determine how to establish the perfect betting line. It is this statistical edge that has contributed to the bottom line of the 50-story lavish casinos and hotels that make up the Las Vegas strip (although sports wagering makes just 3% of gaming revenue).

But now, with sports wagering becoming more commonplace in the American society and the proliferation of available sports data to everyday consumers, there an opportunity to close the gap between casinos and bettors. Similar to how stockbrokers use proprietary projection models to systematically “beat the market”, sports wagering has followed suit.

Recall, a sportsbook’s objective on each bet is to account for an even amount of money wagered on both sides. Often times a betting line is skewed by the inherent biases of an average sports bettor. For example, if the Los Angeles Lakers (a TV market size of over five million people) were to play the San Antonio Spurs (TV market size of just 900 thousand), we might expect a sportsbook to make a line that slightly favors the Spurs because they would expect a disproportionate amount of hometown favorite bets supporting the Lakers. Even the smallest marginal edge, demonstrated by this example, could be enough to be exploited by an adept model. A model, when applied to a large enough dataset, could yield a considerate return on investment.

The goal of this study is to determine if applying machine learning methods to vast sports datasets (specifically within the NBA) can create such a model that would give a bettor this competitive edge over the lines set by a sportsbook.

Chapter 2

The Mathematics of Sports Gambling

References

[1] “Sports Analytics Market Forecasts up to 2024” <https://www.researchandmarkets.com/research/wlpbw8/global_4_5?w=4>

[2] “Curling may finally be having its Moneyball moment”

<https://www.cbc.ca/sports/olympics/winter/curling/curling-moneyball-analytics-1.4623584>

[3] “United States of sports betting: An updated map of where every state stands”

<https://www.espn.com/chalk/story/_/id/19740480/the-united-states-sports-betting-where-all-50-states-stand-legalization>

[4] “Legal Sports Betting Revenue Tracker”

<https://www.legalsportsbetting.com/revenue-tracker/>

[5] “Sports betting market expected to reach $8 billion by 2025”

<https://www.marketwatch.com/story/firms-say-sports-betting-market-to-reach-8-billion-by-2025-2019-11-04>

[6] “NEVADA GAMING CONTROL BOARD Monthly Revenue Report February 2020”

<https://gaming.nv.gov/modules/showdocument.aspx?documentid=16630>