Can Data Science Provide an Edge in Sports Betting?

Doug Marcum
Winter 2020
https://github.com/MarcumDoug/DSC680

Domain

The area to be analyzed for this proposal is the domain of sports betting, in particular professional American football wagering. Specifically, can machine learning be applied in a manner to successfully make predictions that result in an accuracy rate greater than or equal to 55%, the minimum accuracy percentage required for a gambler to maintain profitability.

References and Annotations

Bartels, R. (2019, March 8). Beating the Bookies with Machine Learning. Retrieved from https://www.kdnuggets.com/2019/03/beating-bookies-machine-learning.html.

Richard Bartels, a data scientist with Vantage AI, investigates how to use a custom loss function to identify fair odds, including a detailed example using machine learning to bet on the results of a darts match and how this can assist you in beating the bookmaker.

Browne-Anderson, H. (2019, February 18). Marco and Hugo Discuss the Role of Data Science in Large-Scale Bets and Bookmaking. Retrieved from https://www.datacamp.com/community/blog/data-science-gambling-bookmaking.

An interview with Marco Blume, Trading Director of Pinnacle Sports, and Hugo Browne-Anderson, the host of DataFramed, the DataCamp podcast. The discussion focuses on the role of analytics and data science in bookmaking.

Buchdahl, J. (2018, January 5). The Problem with Data Mining in Sports Betting. Retrieved from https://www.pinnacle.com/en/betting-articles/educational/data-mining-in-sports-betting/4FC29M3VJY59Y4Q3.

Using data as part of a betting strategy is common practice. However, as impressive as some results may appear, the process of producing such results is the important part. The problems with data mining in sports betting are discussed.

Buchdahl, J. (2020, January 10). Using Bayes Factor to Assess Betting Skill. Retrieved from https://www.pinnacle.com/en/betting-articles/educational/bayes-factor-betting-skill-part-one/W9CJUFMSGHQDWKWJ.

Here Joseph Buchdahl explains how the Bayes Factor can be used to test betting skill.

Eulig, S. (2019, September 24). Scraping and Exploring Sports Betting Data – Is Arbitrage Possible? Retrieved from https://towardsdatascience.com/scraping-and-exploring-sports-betting-data-is-arbitrage-possible-a-hands-on-analysis-with-code-2ba656d7f5b.

A deep dive into how one can download live sports betting time series data, parse it, and analyze arbitrage opportunities with Python.

Hubáček, O. & Sourek, G. & Železný, F. (2019, February). Exploiting sports-betting market using machine learning. International Journal of Forecasting. Retrieved from https://www.researchgate.net/publication/331218530_Exploiting_sports-betting_market_using_machine_learning.

An introduction to a forecasting system designed to profit from sports-betting market using machine learning.

Kurcharski, A. (2016, April 30). The Perfect Bet: How the Science of Gambling Influences Everything Around Us. Retrieved from https://www.independent.co.uk/life-style/perfect-bet-how-science-gambling-influences-everything-around-us-a7000741.html.

A look at the book, The Perfect Bet: How science and math are taking the luck out of gambling, and how gamblers are increasingly using scientific ideas to develop successful betting strategies.

Lemmon, B.G. (2019, April 28). A Whole New World: Data Science and Sports Gambling. Retrieved from https://blog.usejournal.com/a-whole-new-world-data-science-and-sports-gambling-ad87dc064162.

A summary of current gambling legislation and how data science methods are being applied to the field.

Sayre, K (2020, September 14). The NFL is Back and Sports Bettors Are Following. Retrieved from https://www.wsj.com/articles/the-nfl-is-back-and-sports-bettors-are-following-11600112703.

A discussion on how the growing legalization of sports wagering in currently 22 states and the lockdowns of Covid-19 have led to increased demand and interest in sports wagering in the NFL, NBA, and MLB.

Stapleton, A. (2020, September 4). The NFL is Now Betting on Once Taboo Gambling Industry. Retrieved from https://www.aol.com/article/news/2020/09/04/the-nfl-is-now-betting-big-on-once-taboo-gambling-industry/24611449/.

An article highlighting the acceptance of gambling by the NFL and public in general by discussing corporate partnerships.

Data

The data set to be utilized was obtained from Stathead.com, a professional sporting statistics resource and subsidiary of Sports Reference. Information pertaining to Sports Reference can be found here: https://www.sports-reference.com/about.html.

With changes to rules of play occurring from decade to decade, the focus of games played will be from the 2000 season through the most current week of play of the 2020 season of the National Football League (NFL). Only regular season games will be reviewed, no preseason or postseason games will be included.

Features will include, but not be limited to, home and away teams, date, game time, the point spread, the over/under of points scored. Additional team offensive and defensive statistics, such as points per game scored and points per game allowed, will be included.

The initial dataset can be found here: https://stathead.com/football/tgl_finder.cgi?request=1&match=game&order_by_asc=1&order_by=vegas_line&year_min=2000&year_max=2020&game_type=R&game_num_min=0&game_num_max=99&week_num_min=1&week_num_max=17&temperature_gtlt=lt.

Example dataset

													Vegas			
Rk	Tm	Year	Date	Time	LTime		Opp	Week	G#	Day	Result	OT	Spread	vs. Line	Over/Under	OU Result
1	PIT	2020	2020-12-02	3:40	3:40		BAL	12	11	Wed	W 19-14			not covered		under
2	BAL	2020	2020-12-02	3:40	3:40	@	PIT	12	11	Wed	L 14-19			not covered		under
3	DEN	2013	2013-10-13	4:05	2:05		JAX	6	6	Sun	W 35-19		-26.5	not covered	53.0	over
4	NWE	2007	2007-11-25	8:23	8:23		PHI	12	11	Sun	W 31-28		-24.5	not covered	51.5	over
5	NWE	2007	2007-12-23	4:15	4:15		MIA	16	15	Sun	W 28-7		-22.5	not covered	45.5	under
6	DAL	2019	2019-09-22	1:00	12:00		MIA	3	3	Sun	W 31-6		-22.0	covered	46.5	under
7	NWE	2007	2007-12-16	1:02	1:02		NYJ	15	14	Sun	W 20-10		-20.5	not covered	41.0	under
8	NWE	2011	2011-12-04	1:02	1:02		IND	13	12	Sun	W 31-24		-20.5	not covered	48.5	over
9	NWE	2019	2019-09-22	1:00	1:00		NYJ	3	3	Sun	W 30-14		-20.5	not covered	43.0	over
10	KAN	2020	2020-11-01	1:00	12:00		NYJ	8	8	Sun	W 35-9		-20.0	covered	49.0	under
11	SEA	2013	2013-09-22	4:26	1:26		JAX	3	3	Sun	W 45-17		-20.0	covered	39.0	over
12	PHI	2002	2002-09-29	1:02	1:02		HOU	4	4	Sun	W 35-17		-19.0	not covered	36.0	over
13	NWE	2007	2007-12-03	8:41	8:41	@	BAL	13	12	Mon	W 27-24		-18.5	not covered	46.5	over
14	IND	2006	2006-10-08	1:02	1:02		TEN	5	5	Sun	W 14-13		-18.0	not covered	47.5	under
15	NWE	2019	2019-09-15	1:00	1:00	@	MIA	2	2	Sun	W 43-0		-18.0	covered	48.5	under
16	OAK	2001	2001-10-07	1:15	1:15		DAL	4	4	Sun	W 28-21		-18.0	not covered	41.5	over
17	STL	2000	2000-10-15	1:02	12:02		ATL	7	6	Sun	W 45-29		-18.0	not covered	58.5	over
18	STL	2001	2001-11-11	1:02	12:02		CAR	9	8	Sun	W 48-14		-18.0	covered	47.0	over
19	NWE	2019	2019-12-29	1:00	1:00		MIA	17	16	Sun	L 24-27		-17.5	not covered	46.0	over
20	BUF	2019	2019-10-20	1:00	1:00		MIA	7	6	Sun	W 31-21		-17.0	not covered	42.5	over

Glossary of terms:

Time	Game Time, Eastern	**Spread**	Vegas Line
LTime	Local Game Time	**vs Line**	Team's W-L-T record against the spread
Week	Week number in season	**Over/Under**	Over/Under
G#	Game number of team	**OU Result**	Was the final score over or under the O/U?

Refinement will need to be conducted to include additional features related to team specific offensive and defensive metrics.

Research Questions and Reason for Analysis

Humans love games, especially games of chance that provide a reward or prize when the correct outcome is selected. We enjoy the highs and lows that come along from engaging in games of chance, gambling. Gambling triggers the brain's reward system which are linked primarily to the pleasure and motivation centers and releases dopamine into the body. This makes the gambler feel elated while they are putting it on the line and taking risks. Dopamine is the dominant power driver and the chief neurotransmitter in the reward system. Gambling stimulates a “thrill” which triggers the reward system to release up to 10 times more than the amount natural rewarding experiences would produce.

With our chemical drive to gamble, and the U.S. Supreme Court ruling, May 2018, in *Murphy v. National Collegiate Athletic Association* that the Professional and Amateur Sports Protection Act violated the 10th Amendment of the U.S. Constitution, thus leading to each state’s ability to legalize sports betting, the industry has witnessed tremendous growth. Publicly traded companies like DraftKings, FanDuel, and Penn Gaming have opened been able to open sportsbooks and mobile sports betting 25 states and the

District of Columbia currently. By the year 2025, sports betting is predicted to generate $8 billion in revenue in the U.S, with some estimates as high as $25 billion.[1]

This industry is unlike the casino games of chance, as the gambler is able to use statistical know how to make a prediction, as opposed to a guess in say a game of roulette. As bookmakers gather more and more data, and play out tens of thousands of potential game scenarios, is it possible for the average better, armed with readily available statistics create a simple model that can provide them an edge in making their selection? This proposal is to determine just that.

Through this proposal and subsequent analysis, it is hoped that the following research questions can be answered:

- What features bear the most importance in determining a winning strategy?
- Is it possible to create a model that is simple enough to allow an average user to engage with and interpret the results without being overly complex?
- Can a model be generated that can make selections with an accuracy rate higher that 55%?

Football, specifically the NFL, was selected for the basis of this proposal as the style of play (multiple players participating separately in offense, defense, and special teams) discourages the effect one player can have on the outcome of each game. Additionally, with a shorter season that Major League Baseball and the National Basketball Association, resting players do not need to be accounted for on a game to game basis.

The data, as with most major sports, is vast and comprehensive. It has been selected allows for a great deal of comparison in determining multiple factors of success.

Methodology

In terms of methodology, the initial steps will be to conduct exploratory data analysis on the data. The gaming industry does not seem to provide the needed data in simple formats, so the data must be gathered and cleaned. Outliers, missing data, and strange distributions will be reviewed and if needed, corrective steps will be taken.

For modeling, multiple classification models will be constructed to determine which path produces the most consistent results, this is for both predicting the potential spread as well as the over/under of each game. The following classifiers will be reviewed: Logistic Regression, Random Forest, K-Nearest Neighbors, and Naïve Bayes. Additionally, polynomial regression will be examined to see if actual scores can be predicted.

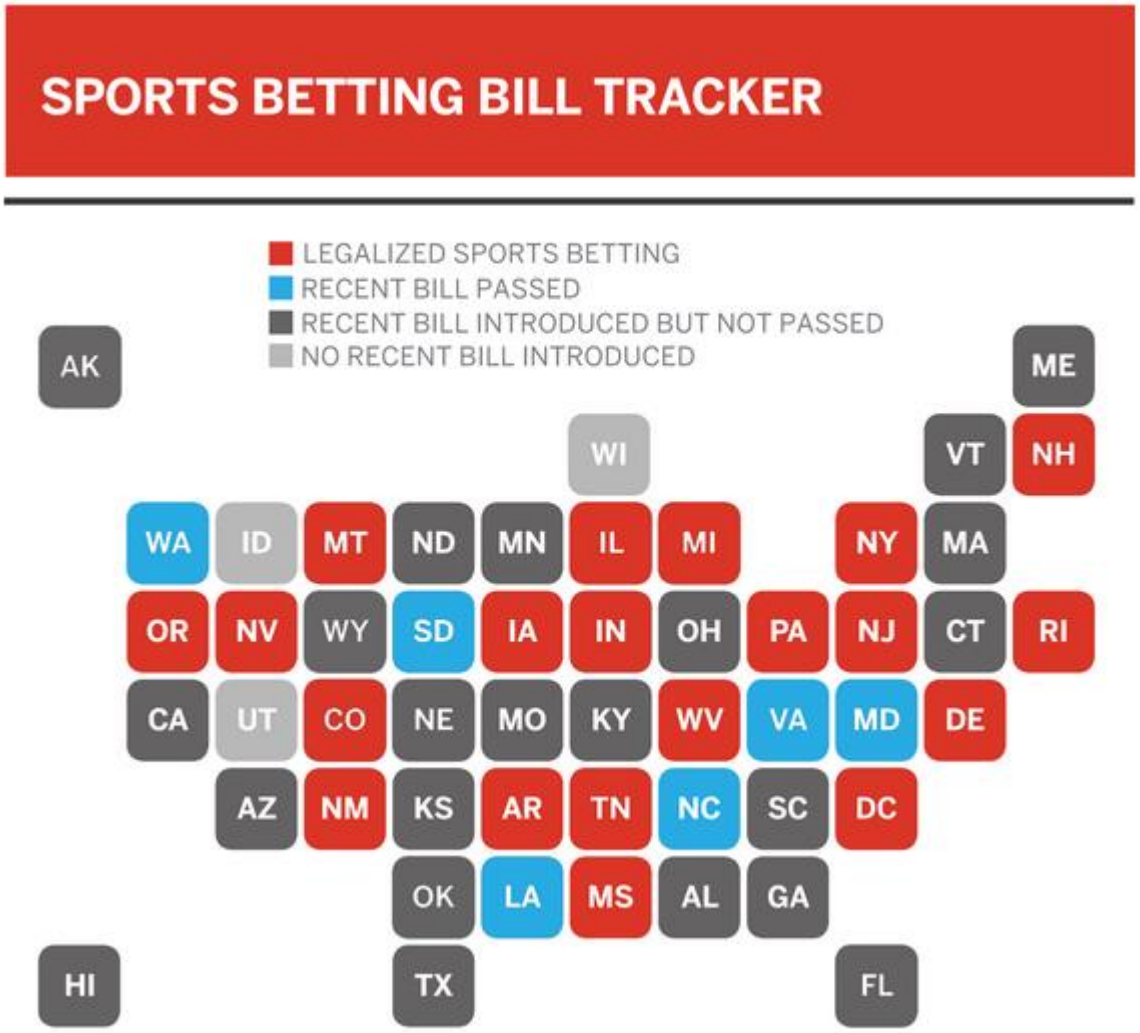
Potential Issues and Challenges

With this project, a few potential issues and challenges may need to be accounted for before diving too deep into the analysis. The main issue is with the number of features that can be taken into consideration. Professional sports account for nearly ever statistical parameter imaginable, so simplifying the initial features need to be committed evaluated and committed. If this is not accounted for, the project could easily stall. In addition, this purpose is to create a simple model that the average better can consume, so overly complicated and advanced features need to be avoided.

[1] Associated Press. (2019, November 4). Sports Betting Market Expected to Reach $8 Billion by 2025. Retrieved from https://www.marketwatch.com/story/firms-say-sports-betting-market-to-reach-8-billion-by-2025-2019-11-04.

One other factor, specific to the 2020 season, is the presence of Covid-19. The virus is a variable that has been causing some interesting issues in the league. For instance, the Denver Broncos were forced to play the New Orleans Saints without a true Quarterback and instead had to utilize a practice squad wide receiver. This case is highly unusual, however, from week to week Covid-19 is causing issues that cannot be accounted for easily.

Conclusion



[2]

Sports betting is not a new phenomenon in the United States, but due to recent legislation, it is experiencing a surge. With 25 states now allowing their residents to wager on sporting events in varying capacities, consumers are searching for tools to assist in making predictions or rather, to gain an edge in their selection process. Resources are available, but often they are not validated, are high priced, and do not provide any true guidance. With the work being conducted under this proposal, the hope is that through data science, the average gambler can be informed in making their selections.

[2] Rodenberg, R. (2020, November 3). United States of Sports Betting: An Updated Map of Where Every State Stands. Retrieved from https://www.espn.com/chalk/story/_/id/19740480/the-united-states-sports-betting-where-all-50-states-stand-legalization.

By analyzing 20 NFL seasons consisting of 256 unique games each season, classification models can be implemented and tested to see if accuracy rates exceeding 55% can consistently be processed. If this is successful, future research into other professional arenas (basketball, baseball, soccer, hockey) can be reviewed. Sports betting, once considered to be taboo and sinful, has grown into a true corporate enterprise. Given this overall popularity, the field is just being explored, and the need for data scientist to engage in this field will continue to increase. More proposals and projects like this will need to be conducted from season to season.