Stan Duel

Beating the Books

Thomas Kummer
Northeastern University
Boston MA USA
kummer.t@northeastern.edu

Harsh Lalwani
Northeastern University
Boston MA USA
lnu.harshl@northeastern.edu

Andrew Enelow
Northeastern University
Boston MA USA
enelow.a@northeastern.edu

Objective

Predict the score for each team in an NFL game to see if it beats the odds of a given sportsbook's over-under and spread.

Background

With the legalization of sports betting in the U.S., betting on games, especially the NFL, has become increasingly popular. There has been no way to "beat the books" through conventional betting strategies, which has caused a large number of gamblers to lose billions of dollars to the sports book companies. This problem is interesting because of how much the sports betting industry has grown over the past few years. Sports betting revenue has grown by more than 15x since 2018 ($920M to $14B). This project introduces a way for people who enjoy gambling to minimize losses by showing the best games to bet on as well as what features affect the outcome of a given game the most.

With an increase in technology, the convenience of sports betting has grown, and along with its legalization there has been a growth in advertising in the industry. This primarily targets young men and gets them to look for a spike in dopamine when betting. When targeting the vulnerable, sports books are able to make millions at the cost of people's livelihood. These problems make it very important to teach the harm of gambling, but that is not always possible. This project hopes to target these people who are being exploited and allow them to make better informed decisions by showing what games are even worth betting on.

Past NFL research has largely concerned predicting game winners, game point spread and wins against the spread.

Bosch (2018): compared neural networks against classical machine-learning methods in predicting game winners. He used data from the 2009 to 2016 seasons and utilized mostly team stats but also average player age, weight, and height. Three types of neural networks were included in his study: traditional ANN (artificial neural networks), LSTM (long short-term memory), and RNN (recurrent neural network).

Beal et al. (2020): compared several machine-learning methods for predicting the winning team for NFL games. They compared nine methods for 1,280 games over five seasons using 42 independent variables that were all team statistics for the current season plus an average for the past season. Much of the past research work for NFL point-spread bets involves classification: beating the spread or not (picking the winning bet given the point spread).

In Bosch's study, the LSTM model achieved 63.1% accuracy, but classical methods like logistic regression slightly outperformed it at 63.33%, followed by support vector machines at 63.25%, and random forests at 62.26%. Despite these advancements, traditional machine learning methods were still unable to consistently beat the NFL betting odds. While some modern machine learning models outperform others, particularly random forests for point-spread prediction, none have consistently overcome the challenges posed by sportsbooks, especially when considering the standard 10% vigorish (the house's cut) on closing odds. However, there is evidence to suggest that modern machine learning methods can perform competitively when using early betting lines (as opposed to the closing lines). Studies also suggest that the historical behavior patterns of individual teams provide more predictive value than patterns derived from all teams collectively.

Evaluation

Our team plans to conduct several experiments to improve prediction accuracy and betting strategies. First, we will focus on data visualization by examining various game parameters, such as stadium location and team rivalries, to detect any anomalies in player performance or game outcomes. For example, we will explore whether certain stadium conditions impact total scores or if specific rivalries lead to more extreme point spreads. Next, we will develop and test multiple machines learning models, including linear regression, random forests, and deep learning, to predict winning bets in both point-spread and over/under categories. By comparing the performance of these models, we aim to identify the most effective approach for predicting game outcomes. Lastly, we will use the models' win probabilities to determine optimal bet sizing for each game, helping us manage risk and potentially maximize returns.

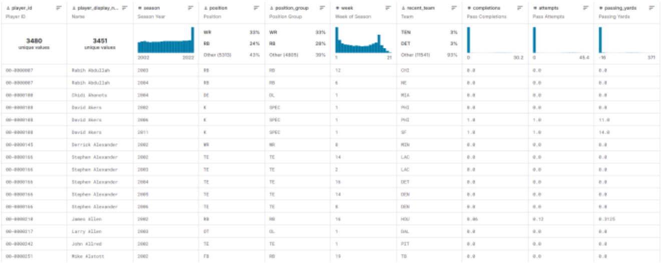
Our project defines success by being able to predict the final scores for each team better than the oddsmakers. This will be a challenge as sportsbooks spend millions of dollars each year perfecting their algorithms to set betting lines, so a more appropriate definition of success would be to achieve over a 50% success rate when setting our preferences for the lines. This would mean that if we were to place a bet on every outcome that we predict, we would end up making a net profit. We can then utalize this success to inform bettors based on our predictions if the payouts for certain lines are good enough to bet on. With a successful model we can also show the top factors that impact on a game, so if bettors don't want to follow the model exactly, they can make better informed decisions when betting.

Data

We have identified two Kaggle datasets for our project. The first contains historical NFL game data, including point spreads, from the 1970s to the present. The second includes player statistics from the 2000s onward. Combining these datasets will help improve score prediction for NFL games.

A key challenge is managing the high dimensionality of the player data, which contains many features, not all of which are relevant. To address this, we will use techniques like correlation analysis, PCA, or feature importance from models like Random Forests to select the most critical features and reduce computational complexity.

These datasets are ideal for this task, offering comprehensive coverage of both game outcomes and player performance. We will treat this as a supervised learning problem, where the target variable is the final score for each team, and the input features include player stats and game-level data (e.g., stadium conditions). The large volume of historical data ensures robust training, validation, and testing for our models.



player_id Player ID	player_display_n... Name	season Season Year	position Position	position_group Position Group	week Week of Season	recent_team Team	completions Pass Completions	attempts Pass Attempts	passing_yards Passing Yards
3480 unique values	3451 unique values	2002 2022	WR 33% RB 24% Other (5313) 43%	WR 33% RB 28% Other (4805) 39%	1 21	TEN 3% DET 3% Other (11541) 93%	0 30.2	0 45.4	-18 371
00-0000007	Rabih Abdullah	2003	RB	RB	12	CHI	0.0	0.0	0.0
00-0000007	Rabih Abdullah	2004	RB	RB	6	NE	0.0	0.0	0.0
00-0000100	Chidi Ahanotu	2004	DE	DL	1	MIA	0.0	0.0	0.0
00-0000108	David Akers	2002	K	SPEC	1	PHI	0.0	0.0	0.0
00-0000108	David Akers	2006	K	SPEC	1	PHI	1.0	1.0	11.0
00-0000108	David Akers	2011	K	SPEC	1	SF	1.0	1.0	14.0
00-0000145	Derrick Alexander	2002	WR	WR	8	MIN	0.0	0.0	0.0
00-0000166	Stephen Alexander	2002	TE	TE	14	LAC	0.0	0.0	0.0
00-0000166	Stephen Alexander	2003	TE	TE	2	LAC	0.0	0.0	0.0
00-0000166	Stephen Alexander	2004	TE	TE	16	DET	0.0	0.0	0.0
00-0000166	Stephen Alexander	2005	TE	TE	14	DEN	0.0	0.0	0.0
00-0000166	Stephen Alexander	2006	TE	TE	8	DEN	0.0	0.0	0.0
00-0000210	James Allen	2002	RB	RB	16	HOU	0.06	0.12	0.3125
00-0000217	Larry Allen	2003	OT	OL	1	DAL	0.0	0.0	0.0
00-0000242	John Allred	2002	TE	TE	1	PIT	0.0	0.0	0.0
00-0000251	Mike Alstott	2002	FB	RB	19	TB	0.0	0.0	0.0

Figure 1: Displays an example of our player data

Timeline

10/03 - 10/21

Clean the data and combine our two datasets into a single usable one to train our models.

10/21 - 11/01

Develop our first model, a simple linear regression model to get a baseline for how it will perform with the data cleaned.

11/01 - 11/15

Develop our final model, a deep learning model using pytorch to make a better prediction than our initial model.

Contributions

Data Cleaning and Preparation: Everyone will assist in cleaning the data and performing feature engineering.

Model Development: Each team member will contribute to the development of the models, with equal responsibility for building both the baseline and deep learning models.

Analysis and Reporting: Team members will work together to analyze the results, draw conclusions, and write up the final report.

REFERENCES

[1] Kaggle NFL Data: NFL Scores and Betting Data

[2] Kaggle NFL Player Data: NFL Player Statistics (2002-present)

[3] Bosch, P., & Bhulai, S. (2018). Predicting the Winner of NFL Games Using Machine and Deep Learning.

[4] Hsu, Y-C. (2021). Using Convolutional Neural Network and Candlestick Representation to Predict Sports Match Outcomes. Applied Sciences, 11(14), 6594. https://doi.org/10.3390/app11146594

[5] Beal, R., Norman, T.J., & Ramchurn, S.D. (2020). A Critical Comparison of Machine Learning Classifiers to Predict Match Outcomes in the NFL. International Journal of Computer Science in Sport, 19(2), 36-50. https://doi.org/10.2478/ijcss-2020-0009

[6] Brandon, D. M. (2024). Predicting NFL Point Spreads via Machine Learning. International Journal of Data Analytics (IJDA), 5(1), 1-18. http://doi.org/10.4018/IJDA.342851