

# Can Specialized ML Beat Sportsbooks in College Football Totals?

Luke Stanton, Alexander Khan

## 1. Introduction

In recent years, sports betting in the United States has grown exponentially along with its legalization. At the same time, public sports data has become more and more accessible, leading to growth in machine-learning (ML) approaches to bet predictions. Most efforts concentrate on high-liquidity markets like moneylines, spreads, or totals in professional leagues like the NFL. In those settings, models can typically only achieve marginal gains over chance once the bookmaker's margin is taken into consideration: under standard -110 pricing, a strategy must exceed approximately 52.38% accuracy to break even. While some models approach or slightly surpass this threshold, their expected value is extremely fragile and highly sensitive to calibration and selection effects. We hypothesize that narrowing the problem to a less modeled market may reveal systematic, exploitable edges that are obscured in broader markets. College football betting markets, especially game totals (over/under points) in niche matchups, could offer a good testing ground. There are over 130 FBS teams with widely varying styles and less media coverage, which makes it much harder for oddsmakers to deploy the same kind of rich, fine-tuned models that they have for the NFL. This creates a good opportunity: if a model can analyze contextual factors and team-specific tendencies more deeply than the sportsbook's baseline approach, it may be able to detect mispriced college totals that the bookmakers and bettors overlook. In this project, we'll focus on college football game totals as our case study. We will build a tailored ML model to predict game scoring outcomes and see if those predictions can identify profitable betting opportunities against the sportsbook's total.

## 2. Research Question

Can a specialized machine learning model identify profitable inefficiencies in sportsbook over/under lines by predicting game scoring outcomes with greater accuracy and better calibration than the market?

## 3. Literature Review

In order to beat sportsbook odds, it's crucial to understand how they are set in the first place. Famously, a 2004 study analyzing bookmaker lines showed that sports books do not "set prices to equalise the amount of money bet on either side of a wager," instead, they "appear to be strategically setting prices in order to exploit bettors' biases" (Levitt, 2004). A lot of these biases have been documented. For instance, a 2022 study found that college football lines "persistently fail to account for censoring bias in the distribution of possible points scored by teams, resulting in exploitable opportunities for profit," showing that the bias "significantly exceeds the typical transaction costs associated with wagers, indicating the market is semi-strong inefficient" (Arscott, 2022). In addition to the structural inefficiencies like censoring bias, researchers have also found behavioral biases in these betting markets. For example, A 2021 literature review outlined that sports betting markets with only two potential outcomes "often produce favorite bias" (Newall and Cortis, 2021). The biases found in these studies are likely more significant in college markets than in the NFL. With much less liquidity, much fewer sophisticated models, and much more variance across hundreds of teams, the same biases that are relatively muted in professional markets could produce systematic inefficiencies in college football.

Meanwhile, the rise of machine learning brings new techniques to sports prediction. A 2024 analysis of ML in sports betting notes that ML techniques have become "essential to managing the complexities of odds setting, risk assessment, and optimization of betting strategy" (Galekwa et al, 2024). Many studies show that these models can reach respectable accuracy levels. For example, a Gaussian process model applied to NFL matchups was able to "successfully pick the game winner over 64% of the time" (Warner, 2010). Even so, translating that accuracy into profit is challenging: that same study noted that its betting scheme fell "short of the mark of 52.4% needed to break even." A key insight is that small advantages can be nullified by variance and the bookmaker's commission, meaning that model calibration and careful selections of bets are crucial. In fact, a 2024 study showed that selecting a model based on calibration "led to profitable betting systems in all cases, with an average ROI of 34.69%," whereas choosing a model based on accuracy "led to an ROI of -35.17% on average" (Walsh and Joshi, 2024). Therefore, a sports betting model needs to correctly estimate confidence along with being accurate.

There is surprisingly little prior scholarly work on college football predictions using machine learning, as most research has focused on the NFL or other professional leagues. One study modeled CFB game winners (South and Egros, 2020), but it did not target betting totals or profitability. Some papers also cover college football betting from a market standpoint (Arscott, 2022), but, to our knowledge, no published studies have applied modern machine learning models to predict college football betting totals for the purpose of beating sportsbooks. Therefore, this project fills a novel gap by extending techniques that have been tried on the NFL to the college domain.

#### 4. Methodology

We plan to train an XGBoost model, which is a decision tree architecture with gradient boosting, to predict the probability that a college football game will go Over (1) or Under (0) the total points line, and then apply an expected value-based decision algorithm to convert those probabilities into positive EV betting opportunities using sportsbook market data. For feature engineering we plan to initialize our model with the rolling past three game averages such as fundamentals like points allowed, time with ball possession, score margins and so forth. We decided on XGBoost for its ability to model non-linear relationships and complex feature interactions in sports relational data, along with its built-in regularization and efficiency on large datasets. The model will secondarily be trained on the residual between the actual total and the sportsbook line to expose and exploit inefficiencies in market pricing. Final betting decisions will be driven by this model output, filtered through a predefined expected value (EV) threshold to ensure only high-value edges are acted upon.

Model performance will be appraised through cross-validation and evaluated using metrics such as AUC and Brier Score. Furthermore, we plan to back-test the entire algorithm on out-of-sample data to estimate ROI and to ensure the robustness and stability of the model's performance across many teams and games. Lastly, we will use the cross-validation and backtesting results for hyperparameter tuning and for plausibly adjusting decision thresholds to optimize predictive accuracy and amplify strong betting signals.

To refine our model further we plan on examining large prediction errors and highest EV wins, essentially the outliers, to identify systemic patterns which could lead us to include pertinent yet less obvious features into our model. From intuition, we may further along the line add to our feature engineering the features of weather and fan attendance. This will further our model into more nuanced and unexplored territory which is novel in principle but may lead us to overfitting problems, so robust CV is necessary to maintain parsimony.

#### 5. Expected Experiments and Datasets

All games, stats, and sportsbook data were sourced via College Football Data API and integrated into Python using the "cfbd-python" package and libraries. The package provides easily accessible structured up to date college football datasets along with built in functions to join tables and preprocess features. We performed some exploratory data analysis and our figures and plots are appended at the end of the document. The visual analyses provide detailed insights into how real game outputs correlate to sportsbook over/under (O/U) lines in FBS regular season games from 2018 to 2025. The initial histogram shows that game totals that score per game broadly have a roughly normal distribution centered between 55 and 60 points. There does exist a strong rightward skew suggesting that a minority of games exceed that of 100-point scores. The subsequent histogram of the distribution of O/U lines provided by bookmakers is significantly tighter and more bell-shaped, with most of its values clustering between 45 and 65 points. A scatter plot that shows a comparison of actual totals to O/U estimates displays a moderate strong meaningful relationship yet withholds considerable variance, especially in the 50–70 point interval. Several of its data points quite substantially appear above or below that of a 45 degree guide line, which suggests that actual total scores often significantly differ from what is anticipated by markets. The vertical spread on the scatter showcases the opportunity for the model to learn and eventually detect residual variance and therefore high EV decisions. Ultimately, we hope that our chosen pre-game features can explain the variance of these residuals and therefore strengthen the signals for positive EV betting.

We plan to conduct assessment experiments of our model using standard benchmarks, including hit rates and ROI as outlined in the literature review. We will initiate variations of our model by adjusting both feature selection and EV decision thresholds. For the XGBoost component specifically, we will fine-tune performance through regularization and depth sweeps experiments. To refine the EV decision threshold, we currently intend to use a greedy brute-force, trial and error approach. To preserve simplicity (and feasibility) and to avoid model drift as odd lines change temporally and different sportsbooks have different odds, we will only use the final pre-game total points line provided through the CollegeFootballData.com API as our single reference point.

#### 6. Timeline

- Week 1: Data Acquisition: pull CFBD games + closing totals/spreads; create data dictionary.
- Week 2: Data Analysis and Feature Engineering: finalize features; lock train/val/test by season; leakage audit.
- Week 3-4: Modeling: baselines; training; pick candidate model family via validation.
- Week 5-6: Tuning and Calibration: hyperparam search; isotonic/Platt calibration; selection rule; limited subgroups
- Week 7: Holdout Evaluation: test on holdout season; metrics; CI vs 52.38%; profit & calibration plots.
- Week 8: Presentation: build deck; speaking plan; rehearsals; advertise; integrate feedback
- Week 9: Report and Extra Credit: write, polish, references; submit.

## **7. References**

Arscott, R. (2022). Market efficiency and censoring bias in college football gambling. SSRN.  
<https://ssrn.com/abstract=4197428>

Galekwa, R. M., Tshimula, J. M., Tajeuna, E. G., & Kyandoghere, K. (2024). A systematic review of machine learning in sports betting: Techniques, challenges, and future directions (arXiv:2410.21484). arXiv.  
<https://arxiv.org/abs/2410.21484>

Levitt, S. D. (2004). Why are gambling markets organised so differently from financial markets? *The Economic Journal*, 114(495), 223–246. <https://doi.org/10.1111/j.1468-0297.2004.00207.x>

Newall, P. W. S., & Cortis, D. (2021). Are sports bettors biased toward longshots, favorites, or both? A literature review. *Risks*, 9(1), 22. <https://doi.org/10.3390/risks9010022>

South, C., & Egros, E. (2020). Forecasting college football game outcomes using modern modeling techniques. *Journal of Sports Analytics*, 6(1), 25–33. <https://doi.org/10.3233/JSA-190314>

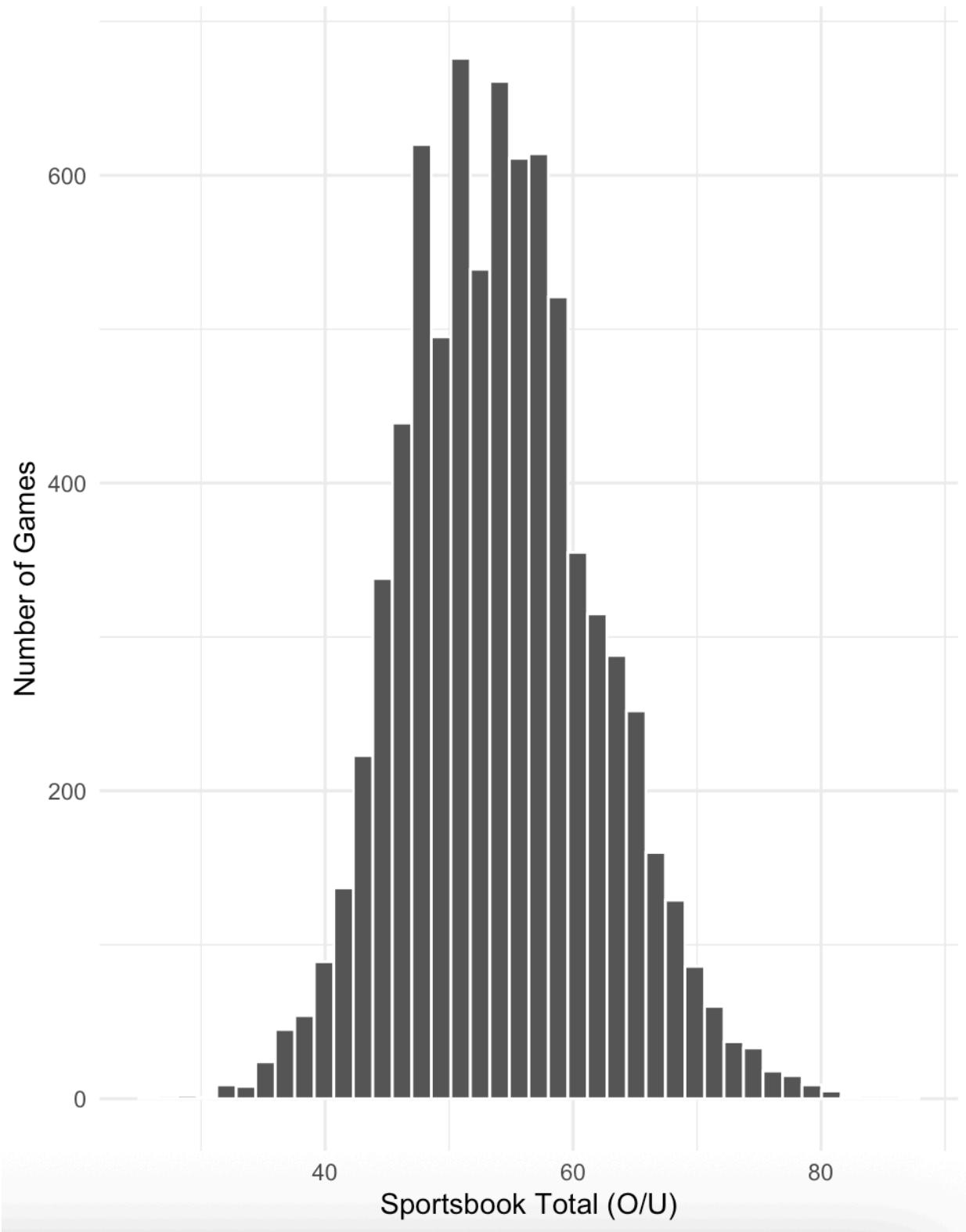
Walsh, C., & Joshi, A. (2023). Machine learning for sports betting: Should model selection be based on accuracy or calibration? (arXiv:2303.06021). arXiv. <https://arxiv.org/abs/2303.06021>

Warner, A. (2010, December 17). Predicting margin of victory in NFL games. Unpublished manuscript, Cornell University. [https://www.cs.cornell.edu/courses/cs6780/2010fa/projects/warner\\_cs6780.pdf](https://www.cs.cornell.edu/courses/cs6780/2010fa/projects/warner_cs6780.pdf)

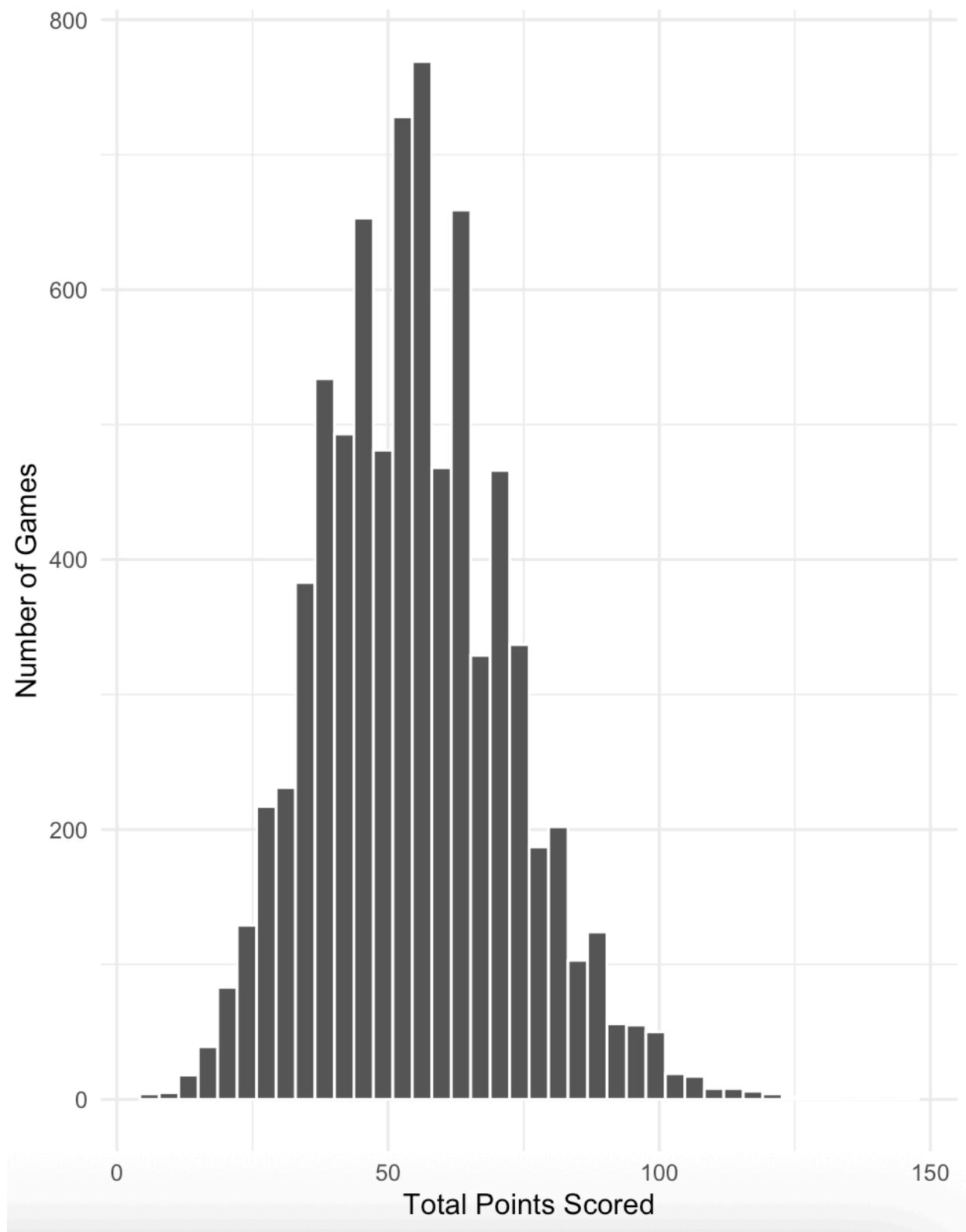
## **8. Appendix**

# Distribution of Sportsbook Over/Under Lines

2018–2025



# Distribution of Actual Total Points (FBS Regular Season) 2018–2025



## Actual vs. Sportsbook Totals

