BetteR Bets: A Decision Support Tool for Sports Betting

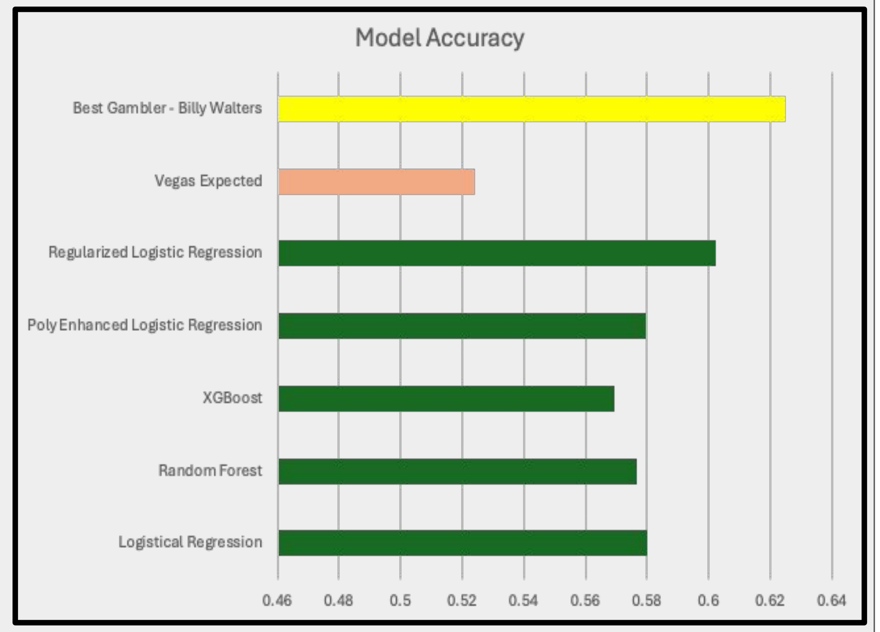
Albert Burton, Russell Cohen, Ryan Katz, Ethan Kobylinski
Matthew A. Lanham
Daniels School of Business
burto143@purdue.edu, cohen100@purdue.edu,
katz30@purdue.edu, ekobylin@purdue.edu, lanhamm@purdue.edu

Abstract

This study designs and develops an R-shiny application that allows end users to train and evaluate various predictive models to estimate the probability of sports outcomes based on historical sports betting data, then allows them to create a betting allocation strategy to maximize expected returns while managing risk. The motivation for this study is that sports betting is a rapidly growing industry, but many bettors rely on intuition rather than data-driven recommendations that can predict outcomes, as well as allocate a betting portfolio based on a user's risk and other preferences. Our tool provides an objective, rather than subjective, approach to betting decisions that leverages machine learning and optimization techniques, which has potential for improved decisions making.

Introduction

This study designs and develops an R-shiny application that allows end users to train and evaluate various predictive models to estimate the probability of sports outcomes based on historical sports betting data, then allows them to create a betting allocation strategy to maximize expected returns while managing risk. The motivation for this study is That sports betting is a Rapidly growing industry, but many bettors rely on intuition rather than data-driven recommendations that can predict outcomes, as wellas allocate a betting portfolio based on a user's risk and other preferences.



Our tool provides an objective, rather than subjective, approach to betting decisions that leverages machine learning and optimization techniques, which has potential for improved decisions making.

Literature Review

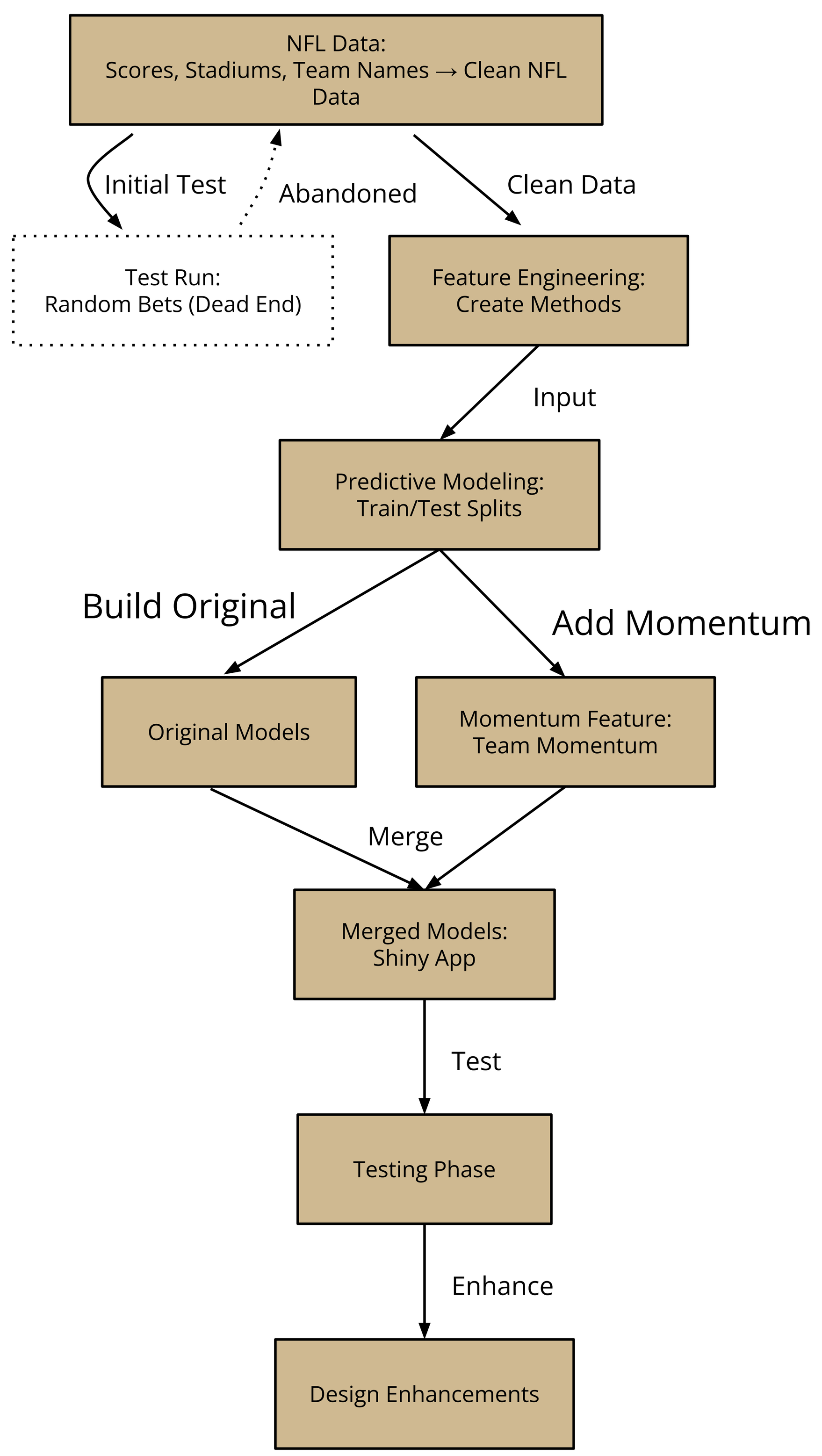
While past studies have explored over/under prediction using logistic regression and neural networks, none have integrated these models into an interactive application using real-time inputs. This project fills that gap by developing a Shiny app that allows users to select matchups and view over/under and spread predictions dynamically.

Study	Data	Model	Application	Accuracy
Matt Gifford and Tuncay Bayrak 2023	Offensive and defensive stats	Decision tree and Logistic regression	Binary, Wins and Losses	81.40%
Kevin Gimple 2006	Offensive, defensive stats & team performance	Logistic Regression	Binary Betting the Spread	~54%
Jim Warner 2010	Offensive, defensive stats + weather	Gaussian process model	Margin of Victory (Spread/ML)	~64%
C. Barry Pfitzner, Steven D. Lang*, and Tracy D. Rishel 2009	Performance stats, last game score, matchup focus	Logistic Regression	Over/Under Lines	54%
Devin Basley, Zachary Strennen, Vinay Maruri, Daven Lagu 2024	Pitching matchups & first-inning batter performance	XGBoost model	NRFI	68%

Methodology

We processed raw NFL data (scores, stadiums, and team names) into a clean dataset. An initial test with random bets was conducted but abandoned. Next, we applied feature engineering techniques to enhance the data before splitting it for predictive modeling.

Multiple models were developed, but we focused on visualizing two key approaches: one using the original dataset and another incorporating a team momentum feature. These models were merged into an interactive Shiny app for further evaluation. The merged models underwent a testing phase, followed by iterative design enhancements to improve performance and usability.

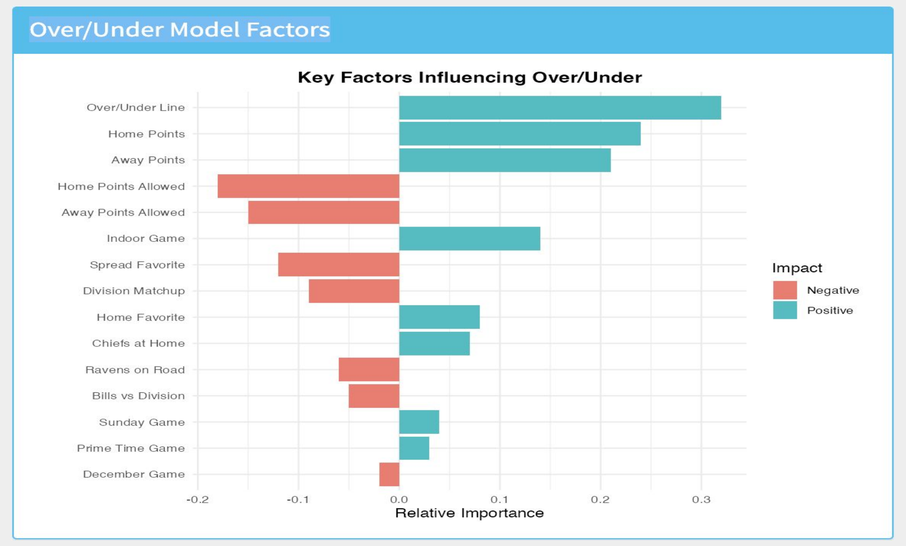


Statistical Results

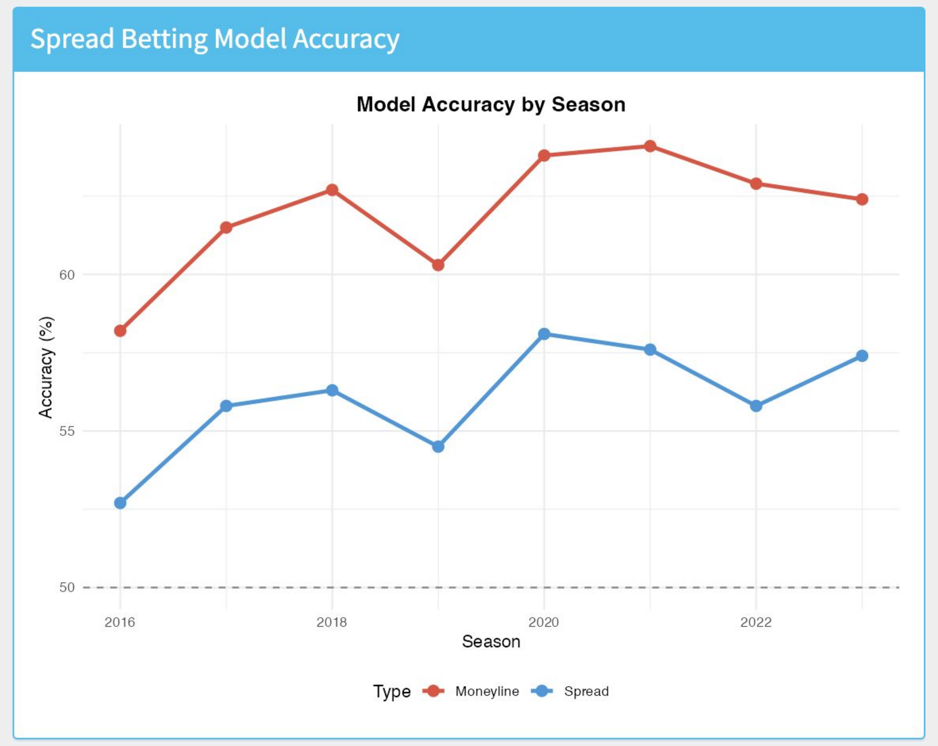
Our best-performing models vary by betting type. Logistic Regression had the highest Over/Under accuracy (0.558), while Poly Enhanced Logistic Regression led in Spread (0.549) and Moneyline (0.612) accuracy (Table 1).

Models	Over/Under Accuracy	Spread Accuracy	Moneyline Accuracy
Logistic Regression	0.558	0.481	0.589
Random Forest	0.542	0.531	0.601
XGBoost	0.551	0.523	0.594
Poly Enhanced LG	0.498	0.549	0.612

As shown in Figure 1, key drivers for Over/Under predictions include the betting line, team scoring, and defensive metrics. Spread accuracy benefits from favored teams and matchup factors.



Moneyline and Spread accuracy (Figure 2) have steadily improved since 2016 due to better feature engineering, including real-time performance metrics and betting line shifts. While Moneyline accuracy remains strong, Spread models show room for improvement.



Expected Impact

Point spreads are intentionally set to be difficult to beat, so even a modest edge—such as a win rate just 5% above random guessing—can be considered a meaningful success. According to experienced professional sports bettors, long-term betting accuracy typically maxes out around 60%.

The table below outlines the expected profit per bet based on the following assumptions: a unit size of $100, odds of -110 for Spread and Over/Under (O/U) wagers, and even odds (+100) for Moneyline (ML) bets. It's important to note that ML odds can vary significantly by game, which may influence the expected value.

In the 2023 NFL season, approximately **73.5 million Americans** engaged in sports betting. If our app offered a **$10/month subscription** and attracted **50,000 paying users annually**, it would generate **$500,000 in revenue**, assuming no operating costs.

From a bettor's perspective, if a user placed **three bets per week** (one each on Spread, ML, and O/U) using our model, their expected **yearly profit** would be **$683.57**. After accounting for the $120 annual subscription fee, the net profit would be **$563.57**—a strong return relative to the investment.

Bet Type	Model Accuracy	EV per $100 Bet	Bets per Season (17 weeks)	Total Expected Profit
Spread	56.30%	$8.23	17	$139.91
Over/Under	55.80%	$7.18	17	$122.06
Moneyline	62.40%	$24.80	17	$421.60
Total	-	-	**51**	**$683.57**

Conclusion/Future Plans

Our models have demonstrated promising accuracy scores, setting a strong foundation for future expansion. In the short term, we're eagerly awaiting the NFL season to test our models in a real-world environment, while concurrently progressing with MLB model tests. Looking ahead, we plan to broaden our scope by exploring opportunities in the NBA and NHL as well. By adapting our approach to these leagues, we aim to validate and refine our models further, ultimately expanding our predictive capabilities across a diverse range of sports.

Acknowledgements

We would like to thank Professor Lanham for his mentorship throughout the project. We also acknowledge the Undergraduate Research Center and Purdue University for providing this valuable opportunity.