Exceeding Expectations in College Football

Lance Pollack

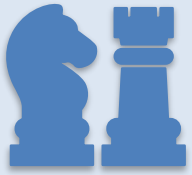
Outline

- Introduction & Background
- Betting Terminology
- Data Collection & Preprocessing
- Feature Selection
- Model Selection
- Evaluation Metrics
- Bootstrapping & Score Multiplier
- Training and Testing Data
- Results
- Feature Importance
- Interpretation of Findings
- Future Work

Introduction

College football is a multi-billion-dollar industry and massive revenue stream for universities.

Revenue is used to fund both athletic and academic programs.



Sports betting has grown into the college football market, adding new layers of complexity to the game.

Betting Terminology

Spread

- handicap to create a balanced betting environment, or make the bets about 50/50
- Florida (-17.5) vs Florida State (+17.5) means Florida is a 17.5 point favorite

Covering the spread

- winning by more or less than the spread
- Florida beat FSU by 20 points and the spread was 17.5, so they "covered" the spread
- If FSU would have lost by less than 17.5 points, they would have covered the spread

Alternate Spread

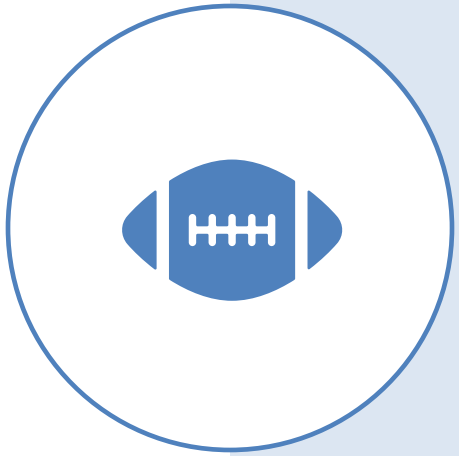
- picking a different spread at odds that are not 50/50
- If I added 13 points to the spread, it would be Florida -30.5 at around 20/80 odds

Book Odds

- odds in the form of –X and +X
- -300 means if you bet $300, you win $100
- +300 means if you bet $100, you win $300 profit

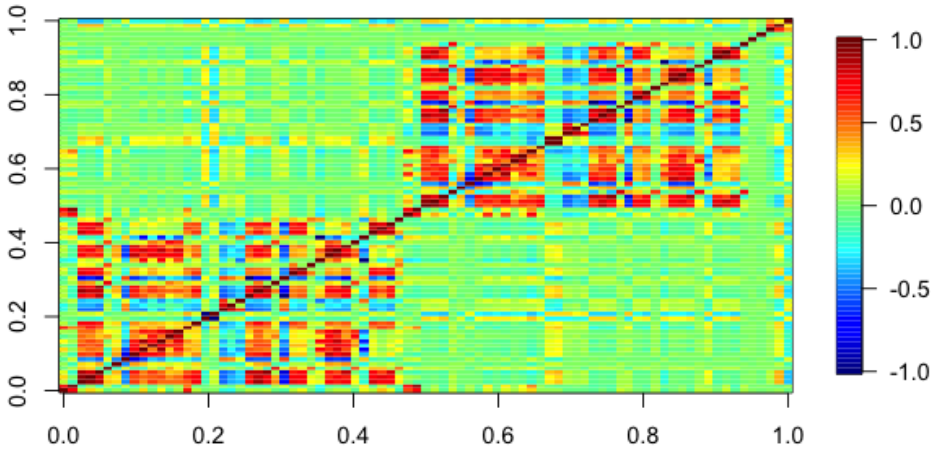
Data Collection and Preprocessing

- Data source: College Football Data
- Span: 2013-2022, Weeks 2-15; approx. 6,000 rows, 80 differential metrics.
- Refined to only week 6-15 to focus on mid to late-season games for stability of metrics.
- Creation of target variables (e.g., alternate line covering labels).



Feature Selection

- Initial predictors: 83 metrics.
- Correlation analysis at 0.80 cutoff reduced predictors to 48 metrics.
- Removed multicollinearity to improve model performance and accelerate run time



Model Selection

Generalized Linear Model (GLM)	GLMnet	Lasso Regression	Ridge Regression
Linear Discriminant Analysis (LDA)	Quadratic Discriminant Analysis (QDA)	K-Nearest Neighbors (KNN)	Support Vector Machine (SVM)
Decision Tree	Extreme Gradient Boosting (XGBoost)	Partial Least Squares (PLS)	

Evaluation Metric

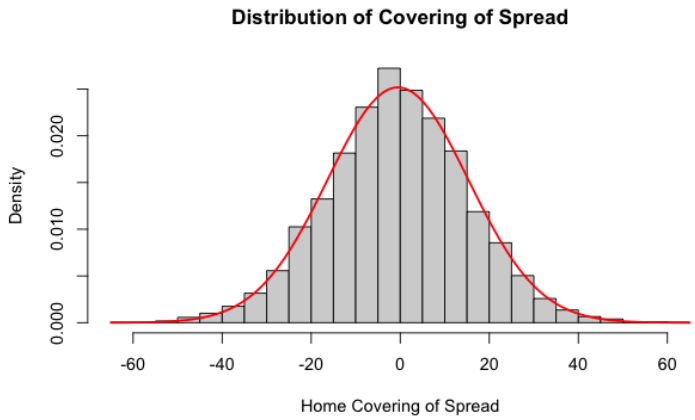
- **Score Metric**:

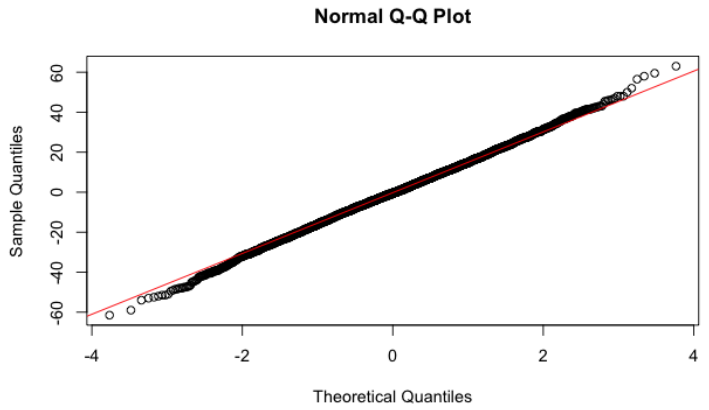
```
(score_multiplier * correct predictions) + (-1 *
incorrect predictions)
```

- *score_multiplier* is the calculated betting odds as a decimal (Ex. +200 becomes 2.0), which will be further discussed in the next section.
- *correct predictions* is the count of when the prediction value was 1 and the actual value was 1, simulating making a bet and winning it.
- *incorrect predictions* is the count of when the prediction value was 1 but the actual value was 0, simulating making a bet and losing it.

Score Multiplier and Bootstrapping

- Probability-Odds conversion:
  - functions `odds_to_prob` and `prob_to_odds` were created to translate between betting odds and probabilities
  - A house edge of 6% (estimated from FanDuel) was added into the conversions
- Parametric Bootstrapping
  - Used to estimate 90% confidence intervals for alternate spreads
  - Assumed normal distribution due to the data's similar structure and distribution

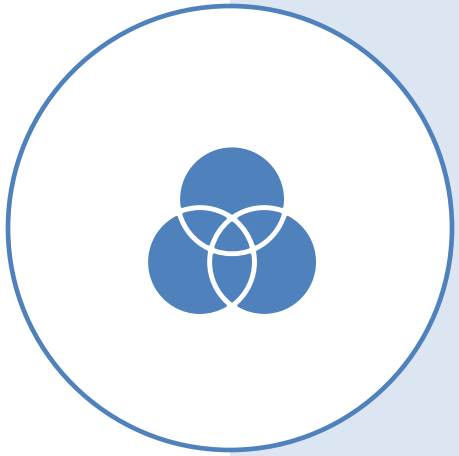




- Score Multiplier for model training
  - The bootstrapping mean value was used as the score multiplier in the score metric for model training
  - The lower CI, mean, and upper CI score multiplier values were all used in model evaluation on testing data
- Ensuring Profitability
  - Use of the confidence interval values allowed for accurate estimation across worst and best case scenarios
  - Confirmed profitability across changing odds

Training and Testing

- Training data: 2013-2019 seasons.
- Testing data: 2020-2022 seasons.
- Cross-validation: 10-fold, repeated 3 times.
- Hyperparameter tuning for optimal performance.
- Away team covering X alternate spread was used as the target variable



Model Performance

Adjusted Line: Spread + 13

Mean Score Multiplier: 3.80

Score Multiplier CI: [3.60, 4.01]

Model	Lower CI Score	Mean Score	Upper CI Score
Gen. Linear Model	20.2	51.2	82.5
GLMnet	39.2	49.8	71.5
Lasso	0.0	0.0	44.7
Ridge	0.0	0.0	44.7
LDA	17.8	47.2	78.5
QDA	4.8	5.4	44.7
K-Nearest Neighbors	32.8	39.5	46.3
SVM (Linear)	0.0	0.0	44.7
Decision Tree	36.4	53.5	73.5
XGBoost	32.4	42.7	78.1
Partial Least Squares	0.6	0.8	44.7

Model Performance

Adjusted Line: Spread + 7

Mean Score Multiplier: 1.97

Score Multiplier CI: [1.88, 2.07]

Model	Lower CI Score	Mean Score	Upper CI Score
Gen. Linear Model	32.1	38.0	44.6
GLMnet	47.0	55.3	64.4
Lasso	29.9	34.5	41.6
Ridge	0.0	0.0	19.0
LDA	33.1	39.0	45.6
QDA	1.3	2.6	19.0
K-Nearest Neighbors	13.1	26.6	52.3
SVM (Linear)	2.8	2.9	19.0
Decision Tree	4.3	11.6	36.4
XGBoost	30.7	45.2	61.3
Partial Least Squares	11.8	12.7	19.0

Model Performance

Adjusted Line: Spread + 3

Mean Score Multiplier: 1.29

Score Multiplier CI: [1.23, 1.34]

Model	Lower CI Score	Mean Score	Upper CI Score
Gen. Linear Model	9.0	9.7	16.5
GLMnet	8.9	10.6	12.2
Lasso	2.5	3.3	22.0
Ridge	9.4	11.0	12.5
LDA	9.0	9.7	15.5
QDA	1.8	5.9	9.8
K-Nearest Neighbors	3.7	8.1	12.2
SVM (Linear)	0.0	0.0	0.0
Decision Tree	12.4	14.0	15.9
XGBoost	6.6	18.4	29.4
Partial Least Squares	4.8	6.2	7.5

Model Performance

Adjusted Line: Spread – 7

Mean Score Multiplier: 0.43

Score Multiplier CI: [0.41, 0.45]

Model	Lower CI Score	Mean Score	Upper CI Score
Gen. Linear Model	3.4	5.7	8.1
GLMnet	6.0	6.8	7.6
Lasso	0.0	0.0	0.0
Ridge	0.0	0.0	0.0
LDA	3.4	5.7	8.1
QDA	0.0	0.0	0.0
K-Nearest Neighbors	0.0	0.0	0.0
SVM (Linear)	0.4	0.4	0.4
Decision Tree	0.0	0.0	0.0
XGBoost	9.4	11.3	13.2
Partial Least Squares	11.8	18.1	24.6

Results Overview

Top-performing models:

- XGBoost
- Decision Tree
- GLMnet

Top performing adjusted lines:

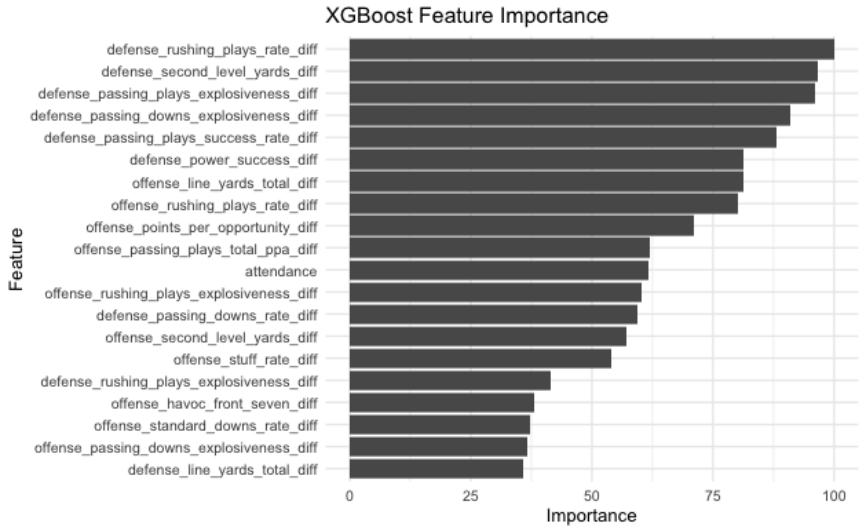
- Spread + 13
- Spread + 7

Profitability:

- The findings suggest there is profitability potential, especially with the best models, as they made upwards of 40 units over 3 season at the lower CI value

Feature Importance

- Top features: Defensive passing plays explosiveness, defensive passing downs rate, offensive stuff rate.
- Defensive metrics dominate as predictors for spread coverage.



Interpretation of Findings

- Ensemble models (e.g., XGBoost, Decision Tree) excelled due to flexibility and ability to handle high-dimensional data
- Models that made strong assumptions about the data performed poorly due to the complexity of the data
- Future work: Incorporate player-level data, real-time metrics, advanced models.

Future Work

Variables:

- Incorporate player-level data like expected points added per player
- Incorporate real-time variables like weather and injuries

Models

- Explore more sophisticated models such as deep learning that excel at handling complex, high-deminsional data
- Further explore hyperparameters of the best models like XGBoost and Decision Trees to improve performance even more

Thank you.
Questions?