Lance Pollack
STA4241
December 9, 2024

Exceeding Expectations in College Football

Introduction

This paper is structured to provide a comprehensive analysis, beginning with an introduction to the business of college football and the fundamentals of sports betting. Following this, the methodology section details the full process and steps from data collection to drawing result. The results section presents the performance of various models. Finally, the conclusion interprets these findings, explores their implications, and outlines potential avenues for future research.

Background

College football acts as a cornerstone of American sports. Annually, it generates billions of dollars, serving as one of the biggest revenue streams for many universities across the country. Schools like the University of Florida leverage their football programs for much more than just athletic success, but also for things like funding academic programs and facility upgrades to enhance their respective school. The immense popularity of college football extends beyond the stadiums and campuses though, permeating the realm of sports betting, where enthusiasts and professionals alike seek to capitalize on their knowledge of the game.

Sports betting, particularly the concept of the spread, plays a pivotal role sports betting due to it leveling the playing field between opposing teams of varying strengths. The spread represents an artificial handicap set by sportsbooks to create a balanced betting environment. For instance, the most recent matchup between Florida and Florida State featured a spread of Florida -17.5, indicating that Florida is expected to win by 17.5 points. Florida won the game by 20 points, so more than the spread. In that case, they are said to have "covered" the spread, offering bettors a basis for wagering. The spread aims to ensure that the likelihood of either team covering is approximately 50/50, although sportsbooks incorporate a house edge to guarantee their profitability.

Importance

The ability to accurately predict whether a team will cover the spread holds significant financial value, given the instantaneous nature of sports betting transactions. While predicting spread coverage is inherently challenging due to the myriad variables influencing game outcomes, it is feasible with the right analytical approaches. This project endeavors to enhance predictive accuracy by not solely relying on the base spread but also exploring the predictability of alternate lines and adjusting for varying odds. Beyond the financial implications for bettors, the insights gleaned from this analysis can be translated into actionable strategies for college football teams and coaches, highlighting critical areas of performance that contribute to dominating games and consistently covering spreads.

Data Origin

The dataset utilized in this study was sourced from College Football Data, encompassing game statistics from the 2013 to 2022 season. The data includes comprehensive metrics for games played from weeks 2 through 15, excluding week 1 due to the absence of meaningful metrics prior to team performance stabilization. The dataset contains about 6000 rows, with each row being a unique team-season-week combination. Using a Python script and the provided API, the dataset was curated by merging home and away team statistics and then calculating the differences in each metric to capture performance disparities between competing teams. This process resulted in approximately 80 differential metrics, alongside variables such as season, week, and team identifiers.

Methodology

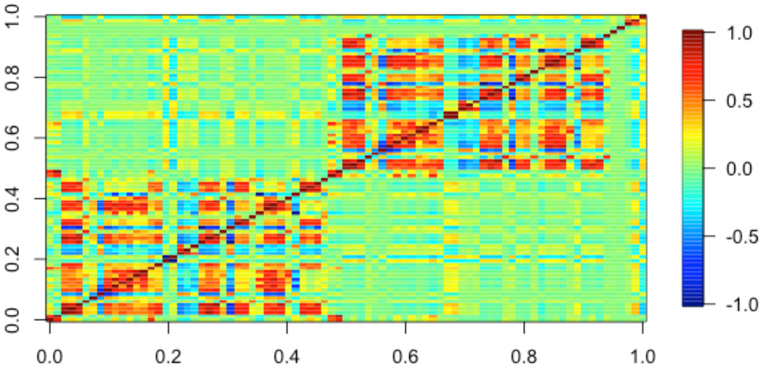
The methodology of this project is meticulously designed to preprocess the dataset, engineer relevant features, and apply a diverse set of machine learning models to predict the likelihood of teams covering the spread in college football games. The following subsections outline each step in detail.

Data Preprocessing

The initial step involved extracting and consolidating data from the College Football Data API using a Python script, as discussed in the data origin section. The approach of getting the differential metrics highlights the strengths and weaknesses of teams compared to their opponents. To ensure the reliability of the metrics, the dataset was refined to include only games from weeks 6 through 15 instead of weeks 2 through 15. Early-season games often involve teams adjusting to new strategies and may include mismatches against lower-division opponents, which can skew performance metrics. By focusing on the mid to late season, the analysis benefits from more stable and representative team performance data. This step also included the creation of label variables to be used as targets for models later on, specifically the home and away alternate line covering label. The labels were assigned 1 when the home/away team covered a designated alternate spread and 0 if not.

Feature Selection

With approximately 80 differential metrics initially present, the next step involved reducing dimensionality to enhance model performance and interpretability. Correlation coefficients were calculated to identify multicollinearity among predictors and then plotted. Red and blue areas indicate high correlation amongst predictors.



Looking at the plot, it is clear there are several predictors very correlated. Features exhibiting a correlation higher than 0.8 or lower than -0.8 with any other feature were deemed redundant and subsequently removed, reducing the number of predictors from 83 to 48. This selection process ensures that the models are not adversely affected by highly correlated variables, which can inflate variance and degrade predictive accuracy. This was also done to accelerate the run time of computationally intense models later on.

Model Selection and Justification

A diverse group of machine learning models was chosen to capture different patterns and relationships within the data. The selected models include:

- **Generalized Linear Models (GLM)**: Provides a baseline for comparison with its interpretability and simplicity.
- **Regularized Models (GLMnet, Lasso, Ridge)**: Helps in handling multicollinearity and enhancing model generalization through regularization techniques.
- **Linear Discriminant Analysis (LDA) and Quadratic Discriminant Analysis (QDA)**: Assumes specific distributions of predictors, useful for understanding class separability.
- **K-Nearest Neighbors (KNN)**: A non-parametric method that captures local patterns in the data.
- **Support Vector Machines (SVM) with Linear Kernel**: Effective in high-dimensional spaces for linearly separable data.
- **Tree-Based Models (Decision Trees, XGBoost)**: Capable of modeling complex interactions and nonlinearities, with ensemble methods enhancing predictive performance.
- **Partial Least Squares (PLS)**: Combines feature extraction and regression, useful for high-dimensional data.

This assortment ensures a comprehensive evaluation of model performances, allowing for the identification of the most effective algorithms for predicting spread coverage. It also allows for a deeper understanding of the data based on the models that perform better or worse.

Evaluation Metrics

To assess the performance of the various models, only one evaluation metric was used:

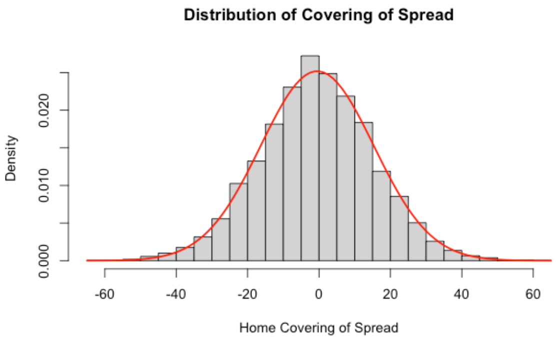
> **Score Metric**: `(score_multiplier * correct predictions) + (-1 * incorrect predictions)`

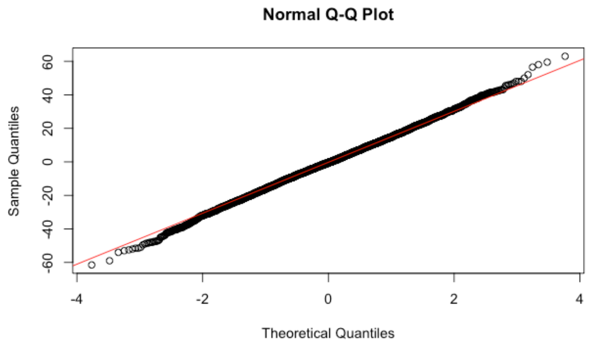
- `score_multiplier` is the calculated betting odds as a decimal (Ex. +200 becomes 2.0), which will be further discussed in the next section.
- `correct predictions` is the count of when the prediction value was 1 and the actual value was 1, simulating making a bet and winning it.
- `incorrect predictions`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

This metric accounts for the financial implications of each prediction based on the score multiplier derived from betting odds. One thing to note is that when the predicted value is 0, then there is no effect to the score metric. The only way to influence the score metric is when the predicted value is 1. This metric was used since the real and only measure of a sports betting model is profitability, which is what the score metric simulates.

Implementation of Score Multiplier and Bootstrapping

A critical aspect of the analysis involved translating betting odds into probabilities and vice versa, factoring in the sportsbook's house edge of approximately 6%, which was estimated using real-world alternate line odds from FanDuel. Functions `odds_to_prob` and `prob_to_odds` were created to facilitate this conversion. To estimate 90% confidence intervals for alternate spread lines, parametric bootstrapping was employed, assuming a normal distribution due to the data's very similar structure to a normal distribution. Below is the density distribution compared to a normal distribution and the normal Q-Q plot to verify the normal distribution assumption:





This approach allowed for the calculation of low and high odds associated with each alternate line, providing a range within which the true odds are likely to fall. The mean value from the bootstrapping was used as the `score_multiplier` in the score metric for model training. The mean value, lower confidence interval value, and upper confidence interval value were all used in evaluating predictions on testing data. Since the exact odds are unknown, it was necessary to be able to accurately estimate low and high odds, as well as test using those odds, to confirm profitability across changing odds

Training and Testing Data Split

The dataset was initially split randomly into training and testing subsets. However, to better simulate real-world predictive scenarios, the data was subsequently partitioned based on seasons. Specifically, data from the 2013 to 2019 seasons was designated as the training set, while data from the 2020 to 2022 seasons served as the testing set. This temporal split ensures that the models are evaluated on their ability to generalize to future, unseen data, reflecting the realistic application of predictive analytics in sports betting.

Model Training and Evaluation

Each selected model was trained on the training dataset and subsequently evaluated on the testing dataset using the score metric. Each model underwent repeated cross validation, specifically a 10-fold cross validation repeated 3 times. Hyperparameter tuning was performed where applicable to optimize model performance. Some models were given pre-defined tuning parameter grids while others were given tuning lengths and tasked with finding optimal tuning parameters. Models were tested with multiple target variables, but ultimately the target used was away alternate line covering. That variable indicated if the away team covered an alternate spread, being 1 if they did and 0 if not. Feature importance was analyzed, particularly for tree-based models like Decision Trees and XGBoost, to identify the most influential predictors in covering spreads by a large margin.

Results

The results section presents the performance of the various machine learning models applied to the dataset, highlighting their effectiveness in predicting whether a team will cover the spread. The evaluation is based on the predefined metrics, providing a comprehensive comparison of each model's strengths and weaknesses, on the away team covering an alternate spread or not. Several alternate spread lines were tested and evaluated.

Model Performance

Model performance was evaluated using the lower confidence interval score multiplier, mean score multiplier, and upper confidence interval score multiplier. Additionally, 20 classification thresholds ranging from 0 to 1 by 0.05 were used to find the max score for each model at each different score multiplier. Each model's performance was measured against the testing dataset (2020-2022 seasons), with their performances summarized in the table below:

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

Using an alternate spread line of 13, most models provide a positive score. While the best score across the different score multipliers varies, there is strong performance from GLM, GLMnet, KNN, Decision Tree and XGBoost models.

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

Model	Lower CI Score	Mean Score	Upper CI Score
Decision Tree	4.3	11.6	36.4
XGBoost	30.7	45.2	61.3
Partial Least Squares	11.8	12.7	19.0

With an adjusted spread of +7, GLMnet and XGBoost continue to perform strongly. GLM, Lasso, and LDA also performed strongly at this spread line.

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

At an adjusted spread of +3, XGBoost and Decision Trees emerge as the top performers, while traditional models like GLM and GLMnet show moderate efficacy. Regularization methods like Ridge also perform decently.

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

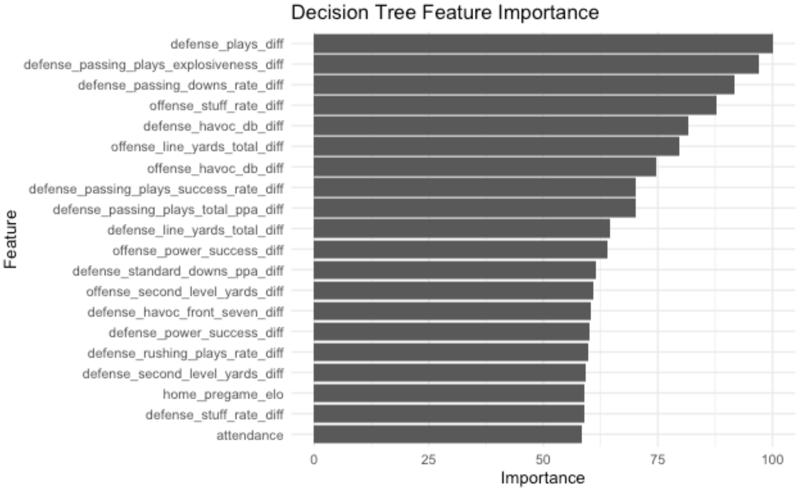
Model	Lower CI Score	Mean Score	Upper CI Score
SVM (Linear)	0.4	0.4	0.4
Decision Tree	0.0	0.0	0.0
XGBoost	9.4	11.3	13.2
Partial Least Squares	11.8	18.1	24.6

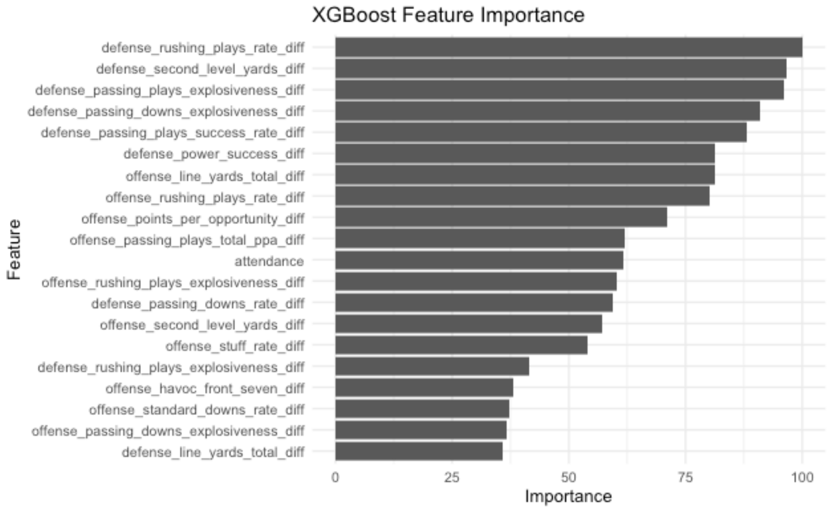
Under a negative adjusted spread of –7, Partial Least Squares and XGBoost remain as the few models showing positive performance. Most other models, including traditional and regularization-based approaches, fail to provide significant scores. The values of 0 indicate the best model was making no predictions at all.

Overall, the best models were GLMnet, Decision Trees, and XGBoost, with XGBoost being the top model. They consistently outperformed the other models across differing spread lines. This is likely due to their flexibility and ability to handle non-linear interactions. This flexibility is needed to for handling complex, high dimensional data such as the dataset used in this project. Some models that struggled in performance across the varying spread lines were simpler models and models that make strong assumptions. Models such as generalized linear models and linear SVMs fail to capture complex patterns within the data, leading to bad performances compared to other more complex models. Additionally, models that make strong assumptions about the relationships in the variables tended to underperform since these assumptions were not held by the data.

Feature Importance

Feature importance was found for the Decision Tree model and the XGBoost model using the alternate spread line of 13, indicating important features for away teams covering the spread by a large margin. The top 20 features for both the Decision Trees and XGBoost models are below:





While the feature lists are different, there are some similarities. Below is a table with features in the top 20 for both decision trees and XGBoost and their respective importance rank in each model:

Feature	Decision Tree Importance Rank	XGBoost Importance Rank
defense_passing_plays_explosiveness_diff	2	3
defense_passing_downs_rate_diff	3	13
offense_stuff_rate_diff	4	15
offense_line_yards_total_diff	6	7
defense_passing_plays_success_rate_diff	8	5
defense_power_success_diff	15	6
defense_second_level_yards_diff	17	2
defense_rushing_plays_rate_diff	16	1
attendance	20	11
defense_line_yards_total_diff	10	20

10 features are present in the top 20 of both models. 3 of the features, are present in the top 10 of both models. One thing to note is there is 7 defensive metrics, 2 offensive metrics, and then the attendance. The inclusion of attendance shows that there is an impact of the crowd that is unrelated to the teams. The inclusion of 7 defensive stats provides insight into the importance of a strong defense, specifically the 2 most important features on average of the difference in defensive passing plays explosiveness, which has the 2nd highest importance for the decision tree model and 3rd for the XGBoost model, and the difference in defensive passing plays success rate, which ranks 8th for the Decision Tree model and 5th for the XGBoost model. In other words, the most important differences are teams that limit big passing plays (explosiveness) and teams that limit passing plays in general. This indicates that teams with strong passing defenses will tend to overperform as away teams, possibly even upsetting some home favorites.

Conclusion

The analysis conducted provides a comprehensive understanding of the factors influencing spread coverage in college football and evaluates the effectiveness of various machine learning models in predicting these outcomes. The following sections interpret the findings, explore their practical implications, and suggest avenues for future research.

Interpretation of Findings

The superior performance of ensemble models, particularly XGBoost and Decision Trees, highlights their proficiency in capturing the intricate relationships and interactions among the diverse set of predictors. These models excel in handling high-dimensional data and mitigating overfitting through techniques like boosting and bagging. The flexibility of those models is what propelled them to outperforming other models that could not handle the high dimensions and complex relationships of the data. Conversely, simpler models like the Generalized Linear Model provided a solid baseline but were outperformed by more complex algorithms, suggesting that the relationships within the data are nonlinear and complex. The relatively lower performance of LDA, QDA, Linear SVM, Lasso, and Ridge models indicates that many assumptions being made by those models do not hold true for the dataset and predictors.

Future Work

Enhanced Features: Future analyses could benefit from incorporating player-level data, advanced metrics like Expected Points Added (EPA), and situational variables such as home-field advantage subtleties. Integrating real-time data, including in-game performance trends and injury reports, could further refine predictive accuracy. Future research should consider integrating external variables like weather conditions, team morale, and coaching strategies to provide a more holistic view of the factors influencing spread coverage. This comprehensive approach can lead to more robust and accurate predictive models.

Advanced Modeling Techniques: Exploring more sophisticated models, such as deep learning architectures or ensemble stacking, could capture even more complex patterns within the data. Further exploration in the best models, specifically XGBoost and Decision Trees, could provide even better performances with certain adjustments to model parameters. Of course, this would be very computationally intensive to find the very best performing model parameters.

Final Thoughts

This study demonstrates the ability of machine learning models, particularly ensemble methods like XGBoost, in predicting the likelihood of college football teams covering alternates spreads. Key performance indicators, such as defensive passing plays explosiveness and defensive passing plays success rate, emerged as pivotal predictors and provided insight into important aspects for outperforming expectations in football. The ability to accurately forecast spread coverage not only holds financial value for sports bettors but also offers actionable insights for athletic programs aiming to enhance team performance. By bridging the gap between sports analytics and betting strategies, this research provides valuable tools for both bettors and athletic departments. The continued advancement in data collection and modeling techniques promises even greater accuracy and utility in future analyses, paving the way for more informed decision-making in the intersection of sports and finance.