Detecting Frauds in College Football

Thuy Linh Luong, Jinghao Meng, Jack Crandell, Nicholas Lambrix

September 2025

Abstract

This project investigates college football team performance, ranking biases, and predictive legitimacy from 2014 to 2025, with the primary goal of developing a "Fraud Score" to identify discrepancies between subjective perceptions and objective on-field results, thereby promoting fairer evaluations in a bias-prone system. Data is sourced from the College Football API and Sports Reference, featuring over 10,000 games with metrics such as scores, ELO ratings, excitement indices, attendance, and aggregated team-season statistics like Win Percentage (WinPct), Margin of Victory (MOV), Simple Rating System (SRS), and Strength of Schedule (SOS). Methods encompass data processing and exploratory data analysis (EDA) using Pandas, NumPy, and Seaborn for visualizations; clustering via KMeans and visualization using PCA to group teams into tiers; Fraud Score computation through threshold-based features and AP Rank; and Random Forest Classifier to evaluate the Fraud Score as well as evaluation for the Random Forest Classifier model itself.

Preliminary results reveal recurring patterns of overrating among historical dominant teams, which support the claims of prestige and regional bias in college football rankings. These findings not only demonstrate the validity of the proposed Fraud Score but also emphasize its potential role in enhancing transparency and fairness.

1 Introduction

1.1 Introduction

College football is one of the most followed sports in the United States, drawing millions of fans, media attention, and significant financial stakes for universities and athletic programs. College football rankings such as the Associated Press (AP) Poll and College Football Playoff (CFP) committee selection serve as the primary influence over evaluating teams and determining post-season opportunities. These rankings can actually do more than reflect team success, they can influence bowl invitations, playoff participation, recruiting power, and institutional prestige.

However, rankings are not strictly based on performance. Rankings often reflect perceptions shaped by brand recognition, historical success, media narratives, and preseason expectations. As a result, there are consistent divergences between how strong a team appears in polls and how well it actually performs according to efficiency metrics and outcomes. This concept of divergence of a team formed the foundation of our study of "fraud teams" - teams whose poll rankings and reputations exceed their true on-field performance.

This project investigated college football teams performance, ranking biases, and predictive legitimacy from 2014 to the present with a primary goal of developing a Fraud Score: a metric to identify discrepancies between subjective perceptions such as the AP polls rankings and CFP selections, and performance-based metrics from the on-field results in order to perform a non-bias evaluation. By creating this score, we seek to quantify perception–performance gaps and test whether certain teams are consistently overrated. For example, debates over Florida State's undefeated 2014 season or TCU's 2022 national championship appearance reflect broader questions about whether polls and the CFP committee accurately identify the "best" teams.

This project also speaks to broader concerns in the literature. Scholars have documented voter bias in the AP Poll [1], structural advantages embedded in the Bowl Championship Series [2], and favoritism toward "power programs" with large fan bases and media profiles. By building a replicable Fraud Score framework, our work contributes to this line of research by offering an applied tool for detecting overrated teams.

1.2 Methodologies in Literature

College football rankings, particularly the Associated Press (AP) poll and the College Football Playoff (CFP) committee, play an important role in determine playoff access, financial allocations, and program reputations. However, decades of scholar research show that these systems are vulnerable to biases and begin to question their reliability. Coleman et al. (2010) documented systematic voter bias in the AP poll, with prestige and regional favoritism influencing rankings more than team quality [1]. Eckard (2011) critiqued the Bowl Championship Series as cartel-like, structurally privileging power conferences over outsiders [2]. Other empirical studies also illuminated these flaws: Witte and Mirabile analyzed AP and Coaches Polls from 2004–2008, uncovering regional favoritism where voters inflate rankings for local teams, leading to inefficiencies in predictive accuracy [3]. Hensley (2015) [4], in a University of California, Berkeley honors thesis, quantified state-level bias in AP ballots, showing voters rank in-state teams ∼2–3 spots higher on average. Stone and Rod (2016) [5] critiqued the CFP, arguing that subjective "eye test" criteria foster inconsistencies. Other research explored private incentives among coaches, uncovering evidence that coaches inflate the rankings of their own conference peers [6]. Collectively, these works establish the ethical and structural rationale for seeking more data-driven, bias-resistant evaluation methods.

Efforts to correct such distortions have produced several objective, quantitative ranking systems. Colley (2002) introduced the Bias-Free Colley Matrix, a method that uses linear algebra to generate team ratings based solely on win-loss outcomes and strength of schedule, avoiding human input entirely [7]. The Simple Rating System (SRS), developed by Drinen (2006) for Pro-Football-Reference [8] and adapted for college by Sports Reference, rates teams via average MOV adjusted for SOS. It solves linear equations: for team i, SRSi = average MOVi + average SRS of opponents, normalized to mean 0 (average team). In college, MOV caps at 24 points curb blowout inflation against FCS foes, and non-D1 games aggregate as one "opponent" for balance. SRS is retrodictive, explaining past results (e.g., correlating highly with win totals) but less predictive due to equal weighting, ignoring recency or injuries [9]. The Elo system, original developed by Elo [10] for chess and adapted to football by Silver [11], provides dynamic ratings.

- Teams start at ∼1500 (average)
- Post-game: $R' = R + K(S - E)$

With K (20–30) for volatility, S as outcome (1/0.5/0), and $E = 1/(1 + 10^{(R_{opp}-R)/400})$.

College adaptations add home-field advantage (HFA) (+55–65 points) and inter-season regression (∼33% toward mean) for turnover [12]. Parallel developments in predictive modeling have produced several objective, quantitative ranking systems. Caballero and Zhao (2024) [13] used Elo for CFP analysis, differing from committee picks, capturing the dynamic evolution of team strength across the season.

In the broader sports analytics field, the push for data-driven alternatives is also rising, coming from sports analytics literature emphasizing quantitative precision over human judgment. Berrar et al. (2019) used K-Means clustering to classify soccer tactics [14], identifying archetypes that enhance outcome prediction. In multisports contexts, Deb et al. (2024) [15] clustered progressive passes in football, while a scoping review by and Fernandez et al. (2024) [16] in the *Electronic Journal of Applied Statistical Analysis* surveyed 278 sports science articles, revealing their ability to categorize play styles, team strengths, and competitive tiers when combined with appropriate validation measures such as silhouette scores. Wang and Huang (2025) clustered tennis matches via deep learning as a proxy for tactical grouping, which is adaptable to college football for performance tiers [17]. Home-field advantage (HFA), a key confounder, is well-documented by Pollard and Pollard (2005) [18] quantified 55–60% home win rates across sports, attributing them to crowd noise and travel. In college football, Risser et al. (2011) [19] used multilevel models to estimate 4–6 point edges varying by venue, while Inan (2020) [20] linked crowd support to outcomes. Boudreaux et al. (2017) [21] noted HFA's decline in replay eras, and Nichols 2014) [22] tied travel to NFL biases, extendable to college.

Ethically, these analytical shortcomings amplify inequities in college athletics, which raised the demands of improving ranking fairness. Edelman (2014) and Eckard (2011) both argue that systemic favoritism and inconsistent ranking criteria exacerbate the competitive imbalance between Power Five and Group of Five schools [23, 2]. In a senior thesis at the University of South Carolina, Oplinger analyzed CFP selections, finding personal biases drive mismatches, exacerbating mental health pressures on athletes [24].

1.3 Project Plan

Building on the literature, our project constructed a Fraud Score that distinguishes between perception-driven rankings and efficiency-driven performance. We focused on the season 2014 to current, coinciding with the CFP era.

We combined College Football Data API [25] with Sports Reference data [26]. Exploratory Data Analysis (EDA) showed visualize relationships between rankings and efficiency to highlight discrepancies. After that we tested methods for detecting fraud teams: Clustering to group teams into elite, contenders, and fraud-prone categories; Creating heuristic thresholds to flag elite vs. borderline performance; Construct a fraud score and conduct supervised learning models to validate the Fraud Score's ability to capture overrated programs.

2 Methods

2.1 Methods: Data Sources, Preparation, and Cleaning

Most of our data came from College Football API [25], which contains data that started in 2014 up until to the present, with a total of 33 columns and 27873 rows. We obtained this data from the College Football Data website which gives us an API where users can download data from the games. We gained access to the College Football Data (CFBD) API using the official python client. Authentication was handled using a Bearer API key, stored as an envrionment variable (CFBD_API_KEY) or entered at runtime when unavailable. We then used this API to download the csv file that is usable for our needs.

ColumnName	Count	Datatype
id	27873	int64
season	27873	int64
week	27873	int64
seasonType	27873	object
startDate	27873	object
startTimeTBD	27873	bool
completed	27873	bool
neutralSite	27873	bool
conferenceGame	27873	bool
venueId	27829	float64
venue	27829	object
homeId	27873	int64
homeTeam	27873	object
homeConference	27796	object
homeClassification	27796	object
homePoints	25023	float64
homeLineScores	25022	object

(a) CFBD Games columns (part 1 of 2)

ColumnName	Count	Datatype
awayId	27873	int64
awayTeam	27873	object
awayConference	27522	object
awayClassification	27522	object
awayPoints	25023	float64
awayLineScores	25022	object
highlights	3391	object
attendance	10038	float64
homePostgameWinProbability	11682	float64
homePregameElo	9539	float64
homePostgameElo	8918	float64
awayPostgameWinProbability	11682	float64
awayPregameElo	9057	float64
awayPostgameElo	8436	float64
excitementIndex	11644	float64
notes	889	object

(b) CFBD Games columns (part 2 of 2)

Figure 1: Description of CFBD Games (2014–2025 API).

The data consists of about 1600 games per year primarily from September to November, since with the start of the 2020 season, the games increased to over 3000 per year due to the addition of lower division games being included. Some of the columns need to be clarified in order to make predictions. The first is total points which is home_points + away_points and the second is margin which is home_points - away_points. Additionally, we created two columns, both for win rate by location, whether the home or away team won. The homeWon column was created by comparing if the homePoints column was higher than the awayPoints column, and storing the value as being an integer with one being a win and zero being a loss. The same was conducted for the awayWon column, except this time comparing if the awayPoints column was greater than the homePoints column. Lastly, we also created an elo difference column which takes homePregameElo - awayPregameElo to calculate a difference in the two.

Our second data source we used is Sports Reference College Football [26]. Sports Reference provides both results stats (wins, losses, points scored) and perception measures (AP poll rankings). It also offers advanced summary ratings: the Simple Rating System (SRS), which adjusts for opponent quality and scoring margins. In this data source, we analyzed a decade of data from 2014 to 2024.

In the Sport Reference Data Source, we used annual team ratings and team standings tables. In order to extract these tables, we used the "Share & Export" feature on Sports Reference website. The feature provided the option of "get table as CSV". The CSV formatted texts were then saved into CSV files. The raw CSV exports contained two-row headers, repeated header rows, and embedded citation text. Several cleaning steps were performed. For the two-row headers such as Scoring – Off/Def, Rushing – Off/Def, we converted them into one-row and changed the name to Scoring_Off, Scoring_Def and relevance. Sports Reference often repeats the header row inside the table so we got these removed as well. Since the data came in separate years, for easier working, we merge all 11 csv files of each ratings or standings into one and added a column of Season.

For the team ratings table, we created derived metrics to enhance comparability across teams and seasons. The Win Percentage (WinPct) is computed as:

$$\text{WinPct} = \frac{\text{Overall\_W}}{\text{Overall\_W} + \text{Overall\_L}}$$

and the Margin of Victory (MOV) is calculated as:

$$\text{MOV} = \text{Scoring\_Off} - \text{Scoring\_Def}$$

For the team standings table, we created two additional metrics. The Conference Strength Differential identifies whether a team struggled within its conference despite padding its record with out-of-conference games, defined as:

$$\text{ConfStrengthDiff} = \text{Conference\_Pct} - \text{Overall\_Pct}$$

The Poll Overperformance / Underperformance Gap, which measures the discrepancy between preseason and final poll rankings, is computed as:

$$\text{PollRankGap} = \text{Polls\_AP\_Pre} - \text{Polls\_AP\_Rank}$$

```
Data columns (total 20 columns):
 #   Column       Non-Null Count  Dtype
---  ------       --------------  -----
 0   Rk           1432 non-null   int64
 1   School       1432 non-null   object
 2   Conf         1432 non-null   object
 3   AP_Rank      275 non-null    float64
 4   Overall_W    1432 non-null   int64
 5   Overall_L    1432 non-null   int64
 6   SRS_OSRS     1430 non-null   float64
 7   SRS_DSRS     1430 non-null   float64
 8   SRS_SRS      1430 non-null   float64
 9   Scoring_Off  1430 non-null   float64
 10  Scoring_Def  1430 non-null   float64
 11  Passing_Off  1430 non-null   float64
 12  Passing_Def  1430 non-null   float64
 13  Rushing_Off  1430 non-null   float64
 14  Rushing_Def  1430 non-null   float64
 15  Total_Off    1430 non-null   float64
 16  Total_Def    1430 non-null   float64
 17  WinPct       1430 non-null   float64
 18  MOV          1430 non-null   float64
 19  Season       1432 non-null   int64
dtypes: float64(14), int64(4), object(2)
```

(a) Ratings

```
Data columns (total 20 columns):
 #   Column               Non-Null Count  Dtype
---  ------               --------------  -----
 0   Rk                   1432 non-null   int64
 1   School               1432 non-null   object
 2   Conf                 1432 non-null   object
 3   Overall_W            1432 non-null   int64
 4   Overall_L            1432 non-null   int64
 5   Overall_Pct          1430 non-null   float64
 6   Conference_W         1378 non-null   float64
 7   Conference_L         1378 non-null   float64
 8   Conference_Pct       1373 non-null   float64
 9   Points_Per_Game_Off  1430 non-null   float64
 10  Points_Per_Game_Def  1430 non-null   float64
 11  SRS_SRS              1430 non-null   float64
 12  SRS_SOS              1430 non-null   float64
 13  Polls_AP_Pre         275 non-null    float64
 14  Polls_AP_High        520 non-null    float64
 15  Polls_AP_Rank        275 non-null    float64
 16  Polls_Notes          5 non-null      object
 17  ConfStrengthDiff     1373 non-null   float64
 18  PollRankGap          157 non-null    float64
 19  Season               1432 non-null   int64
dtypes: float64(13), int64(4), object(3)
```

(b) Standings

Figure 2: Column names, non-null counts, and data types of season 2014 ratings and standings tables from Sports Reference

2.2 Methods: EDA/Visualization

Our first endeavor into exploratory data analysis with this project started with the College Football Data and plotting win rates for home teams vs away teams to see if there is a homefield advantage using a bar plot. We also created a histogram of the excitement index across all games as well as construct a distribution to see how excitement index can vary throughout each season.

We next ventured into the Sports Reference Standings and Ratings and plotted AP Poll ranking against win percentage. We also wanted to examine if there is a correlation between AP Poll rankings and two different measures: win percentage and SRS(Simple Rating System). SRS combines offensive (OSRS) and defensive (DSRS) efficiency while adjusting for strength of schedule, offering a more robust indicator of true team strength than just win-loss records. For each season, we calculated Pearson's correlation between inverted AP Rank (since lower rank = better) and each performance metric. We also plotted PollRankGap against actual win percentage for the first season we had available, 2014.

We also decided that it would be important to see if pregame ELO has any correlation with win rates and to do so we plotted the results for games for both home and away teams with 0 being a loss and 1 being a win. We also applied a logistic regression to this in order to visualize how win rates and pregame ELO are correlated.

2.3 Methods: Clustering Analysis

To identify groups of teams with similar performance and perception levels, we applied K-Means Clustering on the features of each season. Before clustering, we excluded nonnumeric columns such as Team, School, Conference, and Season. All remaining numeric columns were then standardized using z-score scaling. Clustering was performed separately for each season, and we experimented with different values of k (number of clusters) and evaluated cluster quality using the silhouette score and inertia.

After determining the number of clusters using the elbow method (inertia) and silhouette analysis, we generated cluster summaries by computing average values of key performance indicators (SRS, WinPct, MOV, SOS) for each cluster. Using the summaries, we can interpret the clusters in terms of elite performers, solid performers or potential fraud teams, which had strong perception metrics but weaker performance metrics.

We also applied Principal Component Analysis (PCA) to reduce the high-dimensional feature space to two components for visualization. PCA does not affect the clustering itself but allows us to represent multidimensional similarities in a two-dimensional scatterplot. Each point corresponds to a team in the 2019 season, colored by its assigned cluster.

This approach aligns with past literature that emphasizes the gap between human poll rankings and statistical metrics or efficiency metrics in college football rankings [1, 2]. We did the clustering analysis across all numeric features and avoided bias toward teams grouping. This allowed the data to naturally form groups itself.

2.4 Methods: Threshold Construction and Fraud Labeling

The purpose of introducing thresholds is to establish a consistent benchmark for what constitutes "elite" or "borderline" performance across seasons. These thresholds allow for the comparison of teams not only within a given year but also across the broader College Football Playoff (CFP) era. The thresholds form one of the two core components of our fraud score framework.

We derived our thresholds by analyzing the top ten teams in each metric (e.g., SRS, MOV, WinPct) for every season from 2014–2024. For each metric, we calculated the mean performance of those top ten teams and set the threshold slightly above that mean, as shown in Table 1. Teams meeting or exceeding the threshold were flagged with an "EliteHit," while those near the cutoff were marked as "BorderlineHits." A team meeting or exceeding the threshold is therefore performing above the historical average of top-ten teams in that category, approximating what "elite" performance looks like. At the same time, these thresholds help identify teams that may be overrated in polls despite failing to reach elite statistical levels.

Table 1: Key Metrics and Thresholds for Elite/Borderline Classification

Metric	**Elite Threshold**	**Borderline Threshold**
Simple Rating System (SRS)	> 15	$10 \le x \le 15$
Margin of Victory (MOV)	> 20	$15 \le x \le 20$
Win Percentage (WinPct)	> 0.85	$0.75 \le x \le 0.85$
Strength of Schedule (SOS)	> 5	$2 \le x \le 5$
Elo Rating (PostgameElo_End)	> 1700	$1600 \le x \le 1700$
Elo Delta (Season Improvement)	> 100	$50 \le x \le 100$
Total Wins	> 11	$9 \le x \le 11$

Formal Definition. Let each team–season observation be indexed by i, and let $\mathcal{M}$ represent the set of performance metrics (e.g., SRS, MOV, WinPct, SOS, Elo, etc.). For each metric $m \in \mathcal{M}$ with an elite threshold t_m, define the binary indicator

$$\text{EliteHit}_{i,m} = \begin{cases} 1, & \text{if } x_{i,m} \ge t_m \text{ (or } \le t_m \text{ for defensive metrics)}, \\ 0, & \text{otherwise.} \end{cases}$$

The total number of elite thresholds met by team i is then

$$\text{EliteHits}_i = \sum_{m \in \mathcal{M}} \text{EliteHit}_{i,m}.$$

Using both perception and performance, the binary *Fraud Label* is defined as

$$\text{FraudLabel}_i = \begin{cases} 1, & \text{if AP\_High}_i \le 12 \text{ and EliteHits}_i \le 6, \\ 0, & \text{if AP\_High}_i \le 12 \text{ and EliteHits}_i > 6, \\ \text{NA}, & \text{if AP\_High}_i > 12 \text{ or unranked.} \end{cases}$$

Finally, the continuous *Fraud Score* represents the model-estimated probability that a team is fraudulent, scaled to 0–100:

$$\text{FraudScore}_i = 100 \times P(\text{FraudLabel}_i = 1 \mid X_i),$$

where X_i includes only continuous performance metrics (excluding perception and threshold variables). Higher scores indicate a greater likelihood that a highly ranked team underperformed relative to its perceived strength.

To construct the fraud label, we combined perception-based and performance-based indicators. Specifically, the Associated Press Poll served as a proxy for perception through the variable `AP_High`, representing the highest ranking achieved by a team during the season. Teams that at any point reached the AP Top 12 and recorded six or fewer elite thresholds were labeled as *frauds*, while those that also reached the Top 12 but exceeded six elite thresholds were labeled as *contenders*. Teams that were never ranked within the Top 12 were excluded from model training to keep the classification focused on programs that received substantial media and voter attention.

We observed that teams satisfying this fraud criterion—reaching a high poll ranking but failing to achieve elite performance on most objective metrics—consistently fell short of playoff contention or major bowl victories. This empirical validation suggests that the defined thresholds provide a credible basis for distinguishing between perceived and actual team quality.

It is important to note that although the same underlying performance metrics (e.g., SRS, MOV, WinPct) are used both in defining thresholds and as features in our machine learning model, the model does *not* have access to the perception variables (AP rankings) or to the threshold-derived binary indicators (`EliteHit` counts). The model therefore cannot "see" which specific criteria caused a team to fall below the elite benchmark. Instead, it learns generalizable relationships among continuous team-level statistics that best predict the likelihood of a team being classified as a fraud according to our defined framework. This design mitigates, though does not entirely eliminate, the potential for label leakage. While there is some dependency between the data sources, the analysis remains meaningful because the model's role is to validate and interpret which continuous performance factors most strongly align with teams identified as overvalued relative to their poll position.

Finally, because both the thresholds and perception data depend on complete season outcomes, the fraud score is inherently retrospective. Accordingly, the current implementation of the model functions as a *validation tool* rather than a predictive one. Its purpose is to confirm whether the defined thresholds and perception criteria successfully capture the statistical characteristics of overrated teams, not to forecast future cases of fraudulence. Future work could extend this framework to a prospective setting, where early- or mid-season performance metrics are used to predict a team's likelihood of fraudulence before postseason play. Despite this limitation, our retrospective approach provides a transparent and replicable foundation for quantifying discrepancies between perceived and actual team quality in college football.

2.5 Methods: Random Forest Classifier to evaluate Fraud Labeling

To evaluate the Fraud Label we just created for our data, we used Random Forest Classifier model. The model was trained on independent performance-based features such as win percentage, strength of schedule (SOS), SRS, MOV, and Elo metrics. Threshold-derived variables (EliteHits, BorderlineHits) were excluded from the input to avoid circularity. Evaluation was performed using accuracy, F1-score, and ROC-AUC, metrics that balance overall classification performance with the ability to distinguish true "fraud" cases from legitimate contenders. Additionally, feature importance values were conducted as well to identify which variables most strongly influenced the classification.

2.6 Methods: Rationale for Using Random Forest Classifier

We selected the Random Forest Classifier as our primary machine learning model due to its robustness, interpretability, and ability to handle nonlinear relationships among performance metrics.

Random Forests, as an ensemble of decision trees, mitigate overfitting through bootstrap aggregation (bagging), where each tree is trained on a random subset of data and features. This not only improves predictive stability but also ensures generalization across seasons, an essential property for longitudinal sports data. Compared to single decision trees, which tend to overfit, or linear models, which assume additive relationships, Random Forests strike an effective balance between flexibility and reliability.

Another advantage of this approach is its interpretability through feature importance metrics. By quantifying each variable's contribution to model accuracy, we were able to validate that intuitive predictors of team quality—such as Margin of Victory (MOV), Simple Rating System (SRS), and end-of-season Elo—were indeed the most influential. This transparency supports the model's face validity and aligns with existing literature emphasizing the importance of efficiency and dominance metrics in distinguishing elite teams.

Alternative models such as logistic regression and gradient boosting were considered. Logistic regression was ultimately rejected due to its limited ability to model nonlinearities and feature interactions, while gradient boosting (e.g., XGBoost) was deemed less suitable for our project scope, as it required more extensive hyperparameter tuning and introduced interpretability challenges. The Random Forest model, by contrast, provided strong out-of-sample performance with minimal tuning, achieving high accuracy and AUC values while maintaining interpretability—a key goal for ensuring that the "fraud" designation remains explainable and data-driven.

3 Results

3.1 Results: EDA/Visualization

Our first figure resulting from the College Football API is Figure 3 which shows Win Rate by Location as a barplot, whether the winning team was home or away. What it shows us is that there is a clear advantage in being the home team, with the home team having a 52.6% win rate compared to away teams having a 37.0% win rate.

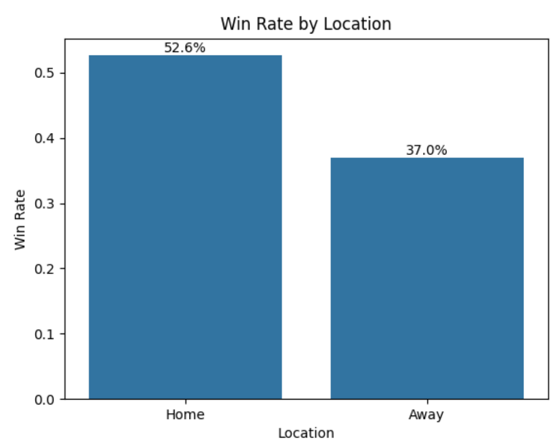


Figure 3: Win Rates by Location(Home vs. Away)

Figure 4 takes the pregame Elos, how strong the team is expected to be before the game and compares it to win rates for both home team and away teams. It also creates a logistic regression line predicting the probability of a team winning given their pregame elo. What it shows is in line with what one might expect for the data, which is that as pregame elo increases so does the winrate. Also notable is that for all elos it appears that the predicted probability of winning for a home team is always greater than that of the away team.

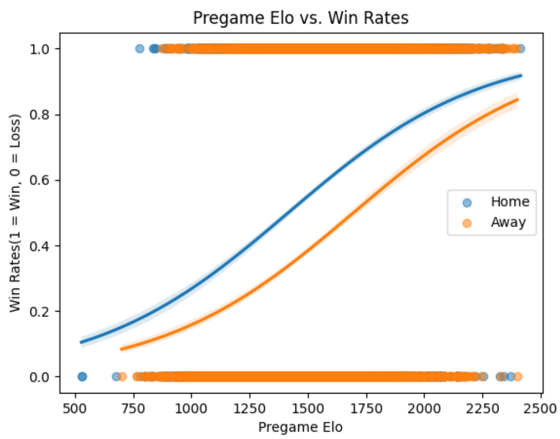


Figure 4: Pregame Elo vs Win Rates

Figure 5 has a similar concept to the previous figure, although this time we took the difference in elo between the home team and away team and compare that to the win rate for just the home team. Again, we created a logistic regression line predicting probability of a home team winning given the difference in their elo compared to the away team. It is also in line with what one might think would happen in the sense that as the elo difference increases so does the win rate and when it is lowest the win rate is lowest.

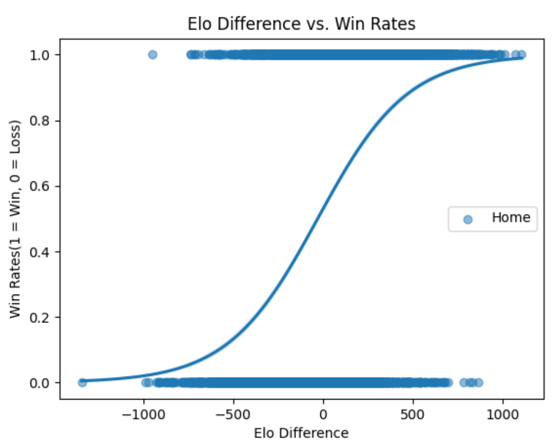


Figure 5: Elo Difference vs Win Rates

The histogram in Figure 6 illustrates the distribution of the Excitement Index across all games in the dataset from 2014 to 2024. The histogram is right-skewed with the majority of games cluster around lower excitement scores. However, the long right tail highlights that a handful subset of games reached very high excitement levels, for example overtime thrillers, rivalry games, and upsets that captured attention nation-wide.

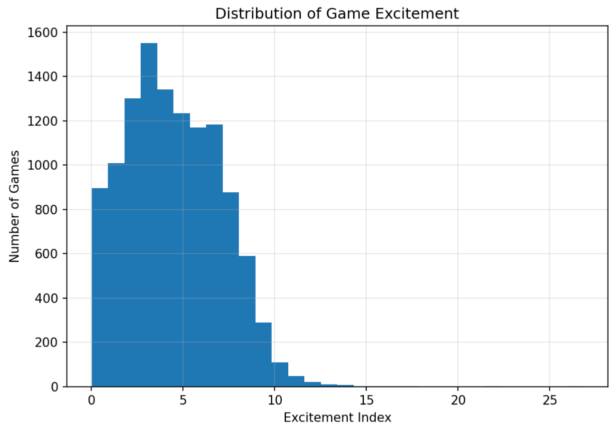


Figure 6: Distribution of the Excitement Index across all games

The boxplot in Figure 7 compares the distribution of the excitement of the game across 2014-2025 seasons. Median values remain fairly stable, but certain years have wider spreads and more frequent high-excitement outliers. Outliers in excitement may help explain why polls overreact or underreact in certain years. If a season produces many high-drama games, poll rankings may reward teams for the hype and popularity rather than consistent efficiency, contributing to fraud risk.

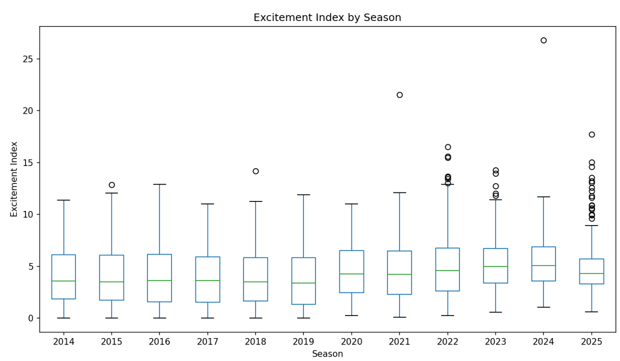


Figure 7: Distribution of game excitement across different seasons

Figure 8 plots the final ranking of the AP Poll against the win percentage. The overall trend shows that stronger ranked teams tend to win more games, but several outliers are noticed. For example, Georgia was highly ranked in the polls despite only a moderate win percentage, suggesting perception exceeded on-field performance. In contrast, Marshall achieved a strong record of Win Percentage despite a lower AP ranking, suggesting a possible underrating. These mismatches highlight where polls can overvalue or undervalue teams and similar patterns which that appear in multiple seasons (2014-2014) can be found in Appendix B. This supports the idea that “fraud teams” are not isolated anomalies but a recurring feature of college football rankings.

Figure 9 shows the correlation between AP Poll rankings and two performance measures: win percentage and SRS (Simple Rating System). SRS combines offensive (OSRS) and defensive (DSRS) efficiency while adjusting for strength of schedule, offering a more robust indicator of true team strength than just win-loss records. For each season, we calculated Pearson’s correlation

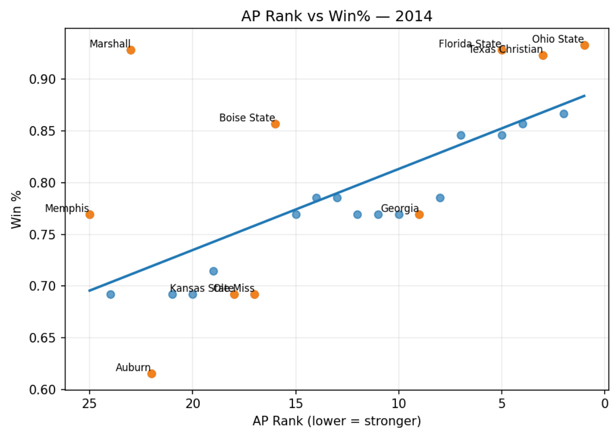


Figure 8: Relationship between final AP Poll ranking and Win Percentage in the 2014 season

between inverted AP Rank (since lower rank = better) and each performance metric. The results indicate that the poll rankings track the win percentage moderately well, but their correlation with SRS is consistently stronger. This suggests that polls reward teams for accumulating wins and are not always overlooking efficiency-based measures.

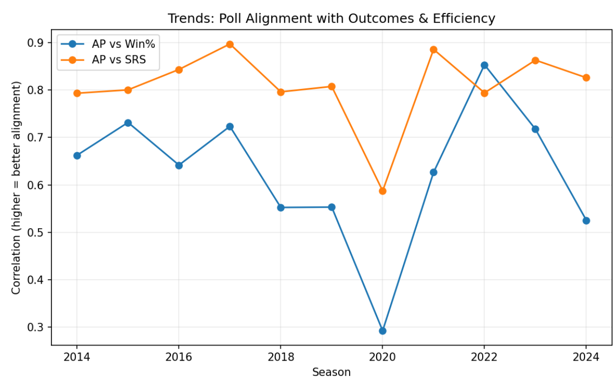


Figure 9: Correlation trend of AP Rank with Win% and SRS (2014–2024)

Figure 10 plots PollRankGap against actual win percentage for the 2014 season. The positive slope indicates that, on average, teams who outperformed preseason polls also won more games. However, significant outliers exist. Auburn and USC illustrate classic "fraud" teams, entering the season with hype but finishing below expectations. In contrast, Oregon, and Ohio State surpassed expectations, demonstrating strong alignment of perception and performance.

3.2 Results: Clustering Analysis

In order to determine the appropriate number of clusters we should use, we evaluated both the silhouette score and inertia across k = 2–10, Figure 11. The silhouette score peaked at k = 2, indicating the strongest separation. However, the solution of 2 clusters might be too simplistic. The inertia curve showed a clear elbow around k = 3 or 4. To balance statistical fit, we selected k = 3 clusters for the main analysis, corresponding to elite, solid, and struggling tiers of teams.

We applied K-Means Clustering Method to the 2019 season with k=3. The result shows three

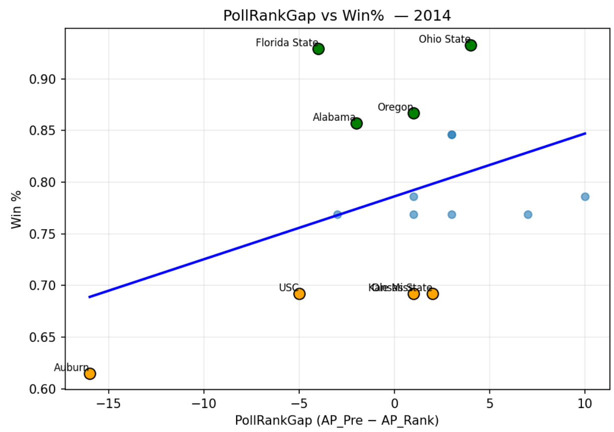
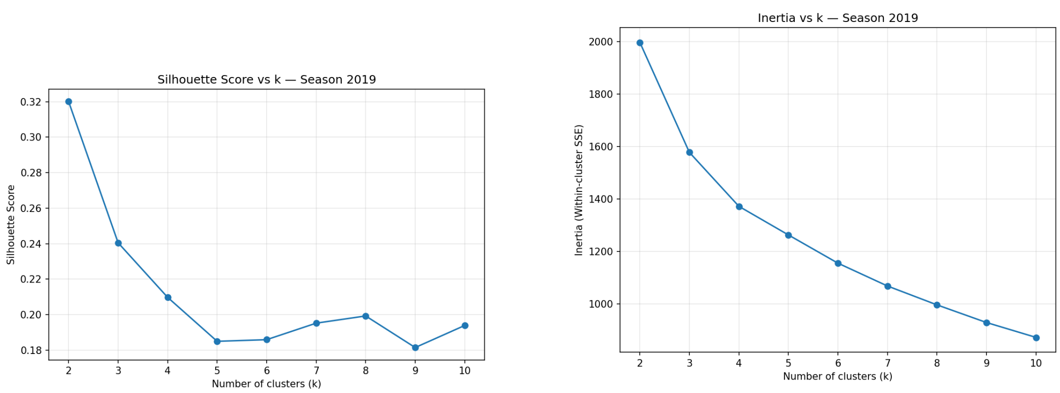


Figure 10: Relationship of Difference between AP Preseason Rank and AP Final Rank and Win Percentage



(a) Silhouette

(b) Inertia

Figure 11: Elbow Line Charts of 2019 Season

clusters with detailed information explained in the following and table 2.

- Cluster 0 (27 teams): This cluster contained the majority of elite teams of 2019. These teams have very high Simple Rating System (SRS) values (mean = 15.15), strong win percentage (mean = 0.79) and dominant Margin of Victory (mean = 50.81).

- Cluster 1 (61 teams): This cluster represented the middle tier teams. The large portion of teams fall into this cluster. Teams in this group achieved moderate success, with an average win percentage (mean = 0.52) and moderate Margin of Victory compare to Cluster 0 (mean = 30.8).

- Cluster 2 (30 teams): This cluster contains struggle or lower tier teams. Teams here had strongly negative SRS values (mean = –11.32) and low win percentages (mean = 0.31). Their performance metrics clearly separated them from both middle tier and elite teams.

The PCA visualization of 2019 clusters, in Figure 12 shows three relatively distinct groupings of teams. Cluster 0 (blue) includes elite teams like Ohio State, Clemson, and LSU, consistent with

Cluster	SRS_SRS	WinPct	MOV	SOS
0	15.148148	0.794438	50.808148	3.542222
1	1.285410	0.515237	30.800328	0.384262
2	-11.318667	0.311966	12.677667	-3.505667

Table 2: Cluster means of key stats in season 2019

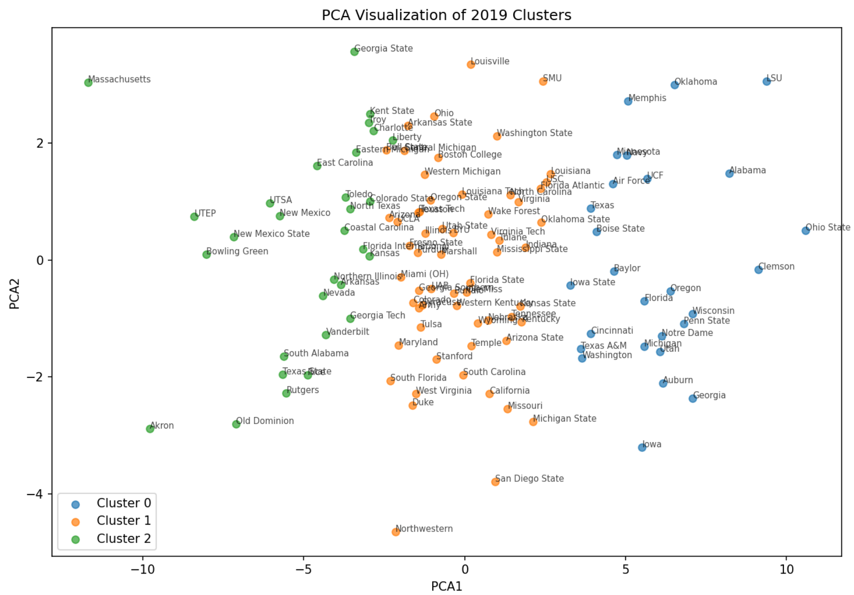


Figure 12: PCA Clustering of 2019 season

their elite performances in both AP polls and efficiency metrics. Cluster 1 (orange) captures middle-tier teams like Michigan State, Missouri, and Arizona State, reflecting strong but not dominant seasons. Cluster 2 (green) consists mainly of struggling or low-performing teams, including Rutgers, UTEP, and Bowling Green, who struggle in both win percentage and efficiency. Although some overlap is visible in the middle clusters, the separation of elite teams from struggling teams is clearly shown.

3.3 Results: Threshold Construction and Fraud Labeling

Using the thresholds defined in Section 2.4, we applied the fraud-labeling framework to all team–seasons from 2014–2024. Teams that reached the Associated Press (AP) Top 12 at any point but registered six or fewer elite threshold hits were labeled as *frauds*, whereas Top 12 teams exceeding six elite hits were labeled as *contenders*. Programs never ranked within the Top 12 were excluded from model training to keep the analysis focused on teams that received significant national attention.

- **Frauds (label = 1):** AP Top 12 teams with EliteHits ≤ 6.
- **Contenders (label = 0):** AP Top 12 teams with EliteHits > 6.
- **Unranked/Irrelevant:** Teams never ranked in the AP Top 12, excluded from modeling.

This framework isolates the subset of teams most subject to perception bias—those widely viewed as championship contenders—and tests whether their on-field statistics supported that perception.

Illustrative examples. Table 3 highlights several 2014 teams labeled as frauds. Each was ranked within the AP Top 12 during the regular season but ultimately failed to reach elite statistical benchmarks or postseason contention.

Table 3: Example Fraud Candidates in 2014, Based on AP Ranking and Elite Hits

Team	AP_High	EliteHits	AP_Rank	FraudLabel
Auburn	2	4	22	1
Mississippi State	1	2	11	1
Georgia Tech	8	5	8	1
Michigan State	5	5	5	1
Wisconsin	11	2	13	1

Aggregate validation. Across all seasons, fraud-labeled teams displayed substantially weaker statistical profiles than contenders (Table 4). On average, contenders recorded over eight more elite hits, a 16 percentage-point higher win rate, and markedly higher efficiency ratings (SRS and MOV). These differences validate that the ≤ 6 EliteHits threshold separates teams of distinct competitive strength.

Table 4: Aggregate Performance Comparison Between Contenders and Frauds (2014–2024)

Group	**EliteHits**	**WinPct**	**SRS**	**MOV**	**Elo_End**
Contenders	10.29	0.863	19.81	57.63	1979
Frauds	2.03	0.703	11.21	43.38	1760

College Football Playoff validation. To evaluate whether these classifications align with real-world postseason outcomes, we examined all fraud-labeled teams that reached the College Football Playoff (CFP) using official CFP historical results [27]. Thirteen such teams appeared between 2014 and 2024, and *every one lost its first playoff game*. The average margin of defeat was 21.5 points, with only two games decided by fewer than two touchdowns (Table 5). By contrast, every national champion and all CFP finalists were labeled as contenders, and all contender–labeled CFP teams either reached the championship game or lost only to another contender. These results confirm that teams failing to reach six elite thresholds rarely overcome their statistical deficiencies when facing truly elite competition.

Table 5: Performance of Fraud-Labeled Teams in the College Football Playoff

Season	Team (Fraud)	Opponent (Contender)	Margin of Defeat
2014	Florida State	Oregon	–39
2015	Oklahoma	Clemson	–20
2015	Michigan State	Alabama	–38
2018	Notre Dame	Clemson	–27
2020	Notre Dame	Alabama	–17
2020	Clemson	Ohio State	–21
2021	Michigan	Georgia	–23
2021	Cincinnati	Alabama	–21
2023	Texas	Washington	–6
2024	Arizona State	Texas	–8
2024	SMU	Penn State	–28
2024	Clemson	Texas	–14
2024	Boise State	Penn State	–17
Average margin of defeat: –21.5 points (n=13)			

Interpretation and limitations. The convergence of statistical and outcome-based evidence supports the validity of the fraud-labeling framework. Fraud teams consistently underperform relative to their perceived strength, while contender teams meet or exceed expectations. The fact that every fraud-labeled CFP participant lost by double-digit margins underscores that the Fraud Score captures genuine performance gaps, not random variance.

Nevertheless, several limitations remain. First, the analysis is inherently *retrospective*: both the thresholds and perception data depend on complete season outcomes, limiting the framework's immediate predictive use. Second, although the model excludes perception variables and threshold flags from its inputs, the underlying performance metrics used to define the label are the same as those used to validate it, introducing a degree of label dependence. Accordingly, the model functions as a *validation tool* that tests the coherence of the fraud-labeling logic rather than a predictive forecasting model. Future work could extend this approach prospectively by using partial-season data or prior-year metrics to predict in-season fraud likelihoods.

Transition to modeling. Taken together, these findings demonstrate that the ≤ 6 EliteHits threshold is both statistically and contextually meaningful, distinguishing true contenders from overrated programs across a decade of college football. The next section evaluates whether a Random Forest classifier trained on continuous team metrics can replicate this classification and identify which performance factors most strongly signal fraudulence.

3.4 Results: Random Forest Classifier Model Performance

The Random Forest Classifier was applied to validate whether continuous team-level statistics could reliably reproduce the fraud labels defined in Section 2.4. The training dataset contained 260 team–seasons, with 192 labeled as frauds and 68 as contenders. This imbalance reflects the frequency with which highly ranked programs fail to sustain championship-level performance. By restricting the sample to ranked teams, the model focuses on distinguishing among perceived contenders rather than contrasting elite and unranked programs.

Overall performance. The classifier achieved strong separation between contenders (label 0) and frauds (label 1). On the holdout test set, accuracy reached 94%, with a precision of 0.949 and recall of 0.974 for fraud cases. The receiver–operating–characteristic area under the curve (ROC–AUC) was 0.987, indicating that the continuous statistics alone captured the patterns that correspond to fraud outcomes. Rather than serving as a predictive forecasting model, these results confirm that the performance metrics used in the Fraud Score framework carry sufficient and consistent signal to reconstruct the retrospective labels.

Season-aware validation. To guard against potential inflation from temporal leakage, we implemented season-aware cross-validation using GroupKFold, ensuring entire seasons were held out of training. Performance remained consistently high across folds, with a mean AUC of 0.982 and mean accuracy of 0.930. This stability demonstrates that the statistical relationships underlying the Fraud Score generalize across seasons rather than reflecting anomalies from individual years.

Feature importance and interpretation. Feature-importance analysis identified the variables most influential in distinguishing frauds from contenders (Table 6). Margin of Victory (MOV) and the Simple Rating System (SRS) were the two dominant predictors, followed by end-of-season Elo ratings and offensive production metrics such as Points For and Total Offense. These results align closely with intuition and with findings in prior sports-analytics literature [1, 2, 5]: true championship-caliber teams consistently win by larger margins and rate higher in efficiency-adjusted systems, while overrated programs often rely on narrower victories, softer schedules, or transient offensive bursts that inflate perception. The Random Forest therefore validates the fraud-labeling scheme by identifying the same statistical foundations—dominance, consistency, and opponent-adjusted efficiency—that historically separate legitimate contenders from teams elevated by prestige or narrative.

Table 6: Top Feature Importances from Random Forest Classifier

Feature	Importance
Margin of Victory (MOV)	0.245
Simple Rating System (SRS_SRS)	0.150
Postgame Elo	0.091
Points For	0.073
Total Offense	0.071
Scoring Offense	0.070
Scoring Defense	0.053
Total Defense	0.043

Relevance and limitations. The Random Forest analysis reinforces the internal validity of the Fraud Score. Its high classification accuracy indicates that the labeling criteria are not arbitrary but correspond to coherent, quantifiable performance patterns. At the same time, several caveats remain. Because the same underlying statistics inform both the fraud label and the model inputs, a degree of label dependence is unavoidable, meaning the model serves primarily as a *validation tool* rather than a forward-looking predictor. The retrospective scope of the data also limits real-time applicability; the model captures what historically characterized overrated teams rather than forecasting future ones. Nevertheless, the consistent feature importance across seasons and the alignment of key predictors with established efficiency metrics suggest that the Fraud Score encapsulates enduring statistical traits of legitimate and illegitimate contenders.

Connection to broader implications. The Random Forest results situate the Fraud Score within a wider discussion on ranking fairness and perception bias in college football. By demonstrating that a small set of objective metrics can explain nearly all variation in perceived team legitimacy, the model underscores both the power and the potential oversight of data-driven evaluation systems.

3.5 Results: Fraud Score Validation

Having established and modeled the Fraud Score, this section evaluates whether the continuous metric behaves consistently with its intended purpose—capturing the gap between perceived and actual performance. Validation was conducted through both statistical and contextual analyses: we examined correlations between the Fraud Score and key performance and perception variables, tested directional relationships using regression, assessed distributional patterns across teams and seasons, and cross-checked high- and low-scoring teams with historical outcomes.

Statistical validation. The correlation heat map in Figure 13 illustrates that the Fraud Score correlates strongly and negatively with efficiency-based performance metrics such as SRS ($r = -0.73$) and Margin of Victory (MOV, $r = -0.73$), and moderately with Win Percentage ($r = -0.53$). Conversely, it shows a positive association with `AP_High` ($r = 0.65$), indicating that teams ranked worse numerically in the polls (larger AP values) tend to receive higher Fraud Scores. This relationship is expected because AP ranking represents perception; its alignment with the Fraud Score confirms internal consistency rather than redundancy. Weaker correlations with contextual variables such as `PollRankGap` and `ConfStrengthDiff` suggest that the Fraud Score more directly captures discrepancies between team efficiency and perception than conference or poll movement effects.

Descriptive statistics further demonstrate the score's separation power. Fraud-labeled teams averaged a Fraud Score of 97.5 ($\sigma = 5.0$), while contenders averaged 8.8 ($\sigma = 8.7$), with negligible overlap between the distributions. The histogram of Fraud Scores (Figure 14) reveals a bimodal pattern—a small cluster near 0 representing legitimate contenders and a large concentration near 100 representing statistically overrated programs—confirming that the continuous measure mirrors the binary distinction defined in Section 2.4.

Regression analysis quantified these relationships. Controlling for AP ranking, each one-point increase in SRS corresponded to an average 1.65-point decrease in Fraud Score ($p < 0.01$), and each additional point of MOV reduced the score by 1.66 ($p < 0.001$), whereas a one-position decline in AP ranking increased the score by approximately 3.0 ($p < 0.001$). Together, SRS, MOV, and AP ranking explained 62% of total variance in the Fraud Score ($R^2 = 0.62$). These results reinforce that higher efficiency systematically lowers a team's likelihood of being labeled fraudulent, while inflated perception raises it.

Temporal consistency. Season-level averages of the Fraud Score ranged from 55.4 to 88.1 (mean ≈ 74), showing that the metric remained stable across the 2014–2024 period rather than being dominated by a single anomalous year. This stability indicates that the score captures enduring relationships between perception and performance rather than season-specific noise.

Illustrative team-level results. To demonstrate how the continuous Fraud Score aligns with key performance indicators, Table 7 reports the top five frauds and two confirmed contenders for three representative years (2014, 2019, and 2022). Consistent with the feature-importance analysis in Section 3.4, teams with near-max Fraud Scores show weaker efficiency metrics—lower MOV,

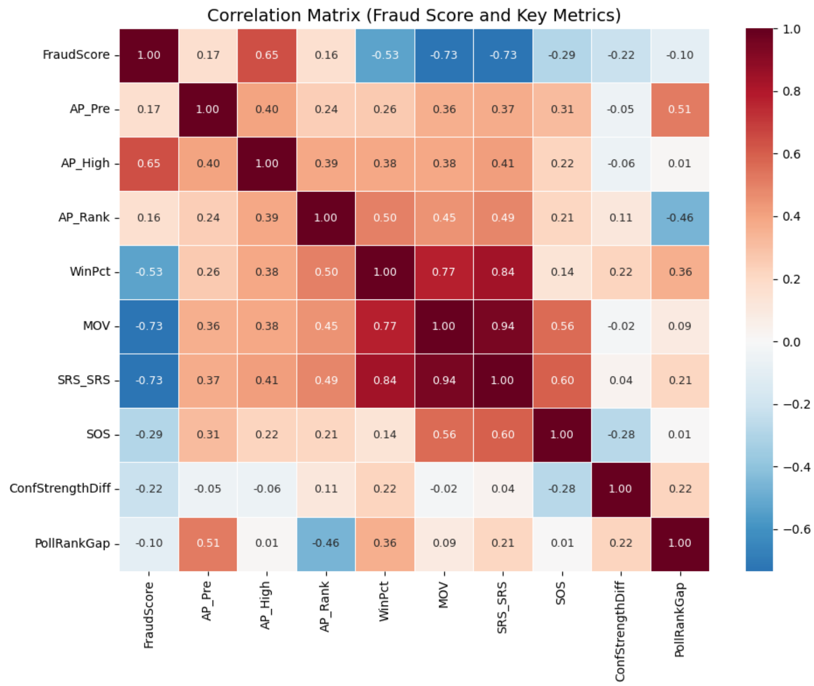


Figure 13: Correlation matrix displaying relationships between the Fraud Score, poll rankings, and performance metrics.

Figure 14: Distribution of Fraud Scores (2014–2024). The right-skewed bimodal shape illustrates clear separation between legitimate contenders and statistically overrated teams.

SRS, and Elo ratings—despite strong poll rankings. In contrast, legitimate contenders exhibit dominant margins, higher SRS, and stronger Elo ratings, producing single-digit Fraud Scores.

Table 7: Fraud Scores and Key Performance Metrics for Selected Teams (2014, 2019, 2022)

Season	Team	FraudScore	AP High	MOV	SRS	Elo End	WinPct
2014	Mississippi State	100.0	1	15.0	15.0	1820	0.77
2014	Texas A&M	100.0	6	10.0	9.9	1790	0.62
2014	Alabama	5.2	1	20.3	20.3	1990	0.86
2019	Texas A&M	100.0	12	10.0	10.0	1775	0.62
2019	LSU	1.8	1	25.8	25.8	2050	1.00
2022	Clemson	100.0	4	13.6	13.6	1840	0.79
2022	Michigan	4.0	2	20.6	20.6	2020	0.93

High Fraud Scores consistently coincide with weaker efficiency metrics and moderate win percentages despite high AP rankings, while low Fraud Scores belong to true contenders whose dominance mirrors the model's most influential predictors (MOV, SRS, and Elo). This reinforces that the Fraud Score's scaling is consistent with the statistical foundations of the Random Forest model.

Case validation. To confirm external validity, we compared model-predicted Fraud Scores with historical outcomes from official College Football Playoff (CFP) archives [27], ESPN, and Sports Reference. Teams receiving Fraud Scores near 100 consistently underperformed relative to their poll rankings. For instance, **Mississippi State (2014)** was assigned a Fraud Score of 100 despite reaching #1 in the AP Poll midseason; the Bulldogs finished 10–3 and lost 49–34 to Georgia Tech in the Orange Bowl. Similar patterns appeared for **Texas A&M (2019)** and **Clemson (2022)**, which both posted high Fraud Scores and failed to meet their championship expectations. These outcomes reinforce that elevated Fraud Scores correspond to tangible postseason underperformance, confirming the metric's face validity.

Interpretation and limitations. The Fraud Score behaves in line with theoretical expectations: it declines as efficiency and dominance rise, increases with inflated perception, and remains stable across time. While these validations rely on retrospective data and share some variables with the original labeling framework, their directional consistency and explanatory power affirm that the Fraud Score captures a coherent and reproducible construct. The analysis therefore substantiates the score's use as a diagnostic measure of perception–performance discrepancies rather than a forward-looking predictor. Future extensions could explore predictive calibration by evaluating partial-season Fraud Scores or applying the framework prospectively to upcoming seasons.

Synthesis. Together, the statistical, temporal, and contextual validations demonstrate that the Fraud Score is internally consistent, externally grounded, and temporally stable. It reliably quantifies when team perception exceeds on-field production, offering a transparent, data-driven complement to subjective ranking systems in college football.

4 Discussion

4.1 Discussion: EDA/Visualization

Our exploratory data analysis provided foundational insights into the relationship between perception-based measures, such as AP Poll rankings, and performance-based metrics, including Win Percentage (WinPct), Simple Rating System (SRS), and Margin of Victory (MOV). Most notably, the figures we created depicting the relationship between final AP Poll ranking and win percentage for each season was helpful in identifying "frauds" early on in the project. The EDA also helped identify consistent patterns across seasons. For example, teams such as Oklahoma, Clemson, Ohio State frequently appeared near the top of both AP and performance metrics, while others, like UCF or Baylor in certain years, achieved strong records but weaker efficiency scores, consistent with the perception bias documented [1, 5]. This evidence guided the later design of our Fraud Score, which distinguishes between perception and performance to determine these disparities.

4.2 Discussion: Clustering Analysis

The PCA clustering of the 2019 season in Figure 12, when viewed alongside the AP Preseason and Final Rankings in Table 8, highlights both alignment and divergence between perception and performance. Elite teams such as LSU, Ohio State, and Clemson not only clustered in the high-performance group but also finished within the top of the AP rankings. These teams serve as an example for what a "true contender" looks like in our framework.

However, several divergences exist when comparing AP ranks to cluster placement. For instance, teams such as Texas, Michigan and Texas A&M entered the season highly ranked in the AP poll but were placed close to the mid-tier clusters in the PCA analysis as well as dropping to lower AP Rank at the end of the season. This gap emphasizes the issue of voter bias in preseason polls identified by Coleman et al. (2010), where historical prestige or media may inflate early-season expectations [1].

Table 8: AP Poll Rankings: Preseason, Highest, and Final (2019 Season)

School	AP Pre	AP High	AP Rank
Clemson	1	1	2
Alabama	2	1	8
Georgia	3	3	4
Oklahoma	4	4	7
Ohio State	5	2	3
LSU	6	1	1
Michigan	7	7	18
Florida	8	6	6
Notre Dame	9	7	12
Texas	10	9	25
Oregon	11	5	5
Texas A&M	12	12	–
Washington	13	13	–
Utah	14	5	16
Penn State	15	5	9
Auburn	16	7	14

School	AP Pre	AP High	AP Rank
UCF	17	15	24
Michigan State	18	18	–
Wisconsin	19	6	11
Iowa	20	14	15
Iowa State	21	21	–
Syracuse	22	21	–
Washington State	23	19	–
Nebraska	24	24	–
Stanford	25	23	–
Wake Forest	–	19	–
Virginia	–	18	–
Virginia Tech	–	23	–
Cincinnati	–	17	21
Memphis	–	15	17
Navy	–	20	20
SMU	–	15	–
Baylor	–	8	13
Kansas State	–	20	–
Oklahoma State	–	21	–
Texas Christian	–	25	–
Indiana	–	24	–
Maryland	–	21	–
Minnesota	–	7	10
Boise State	–	14	23
Air Force	–	22	22
San Diego State	–	24	–
California	–	15	–
USC	–	22	–
Arizona State	–	17	–
Missouri	–	22	–
Appalachian State	–	19	19

4.3 Discussion: Fraud Score

The correlation matrix demonstrates that Fraud Score strongly negatively correlated with SRS and MOV and moderately negatively correlated with WinPct, indicating that teams with weaker efficiency and smaller margins of victory receive higher (worse) Fraud Scores. These results mirror prior findings that SRS and Elo are better in capture team quality than human polls [8, 11]. Consistent with the literature examples of poll overvaluation, teams like Mississippi State (2014) and Clemson (2022) were identified as high-fraud cases, where both achieved early top-5 rankings but failed to maintain elite performance. These instances align with Coleman et al. (2010) and Stone and Rod (2016), who described how early prestige and the "eye test" can sustain inflated perceptions even after on-field evidence diverges [1, 5].

4.4 Discussion: Limitations and Future Refinement

Although the Fraud Score helps identify teams that might be overrated or underrated, it still inherits limitations from both its design and data sources. Since AP-based perception variables contribute directly to the fraud labeling. The correlations between poll variables and fraud score in the correlation matrix reflect shared construction rather than causal relationships.

Second, the thresholds we used to define "elite" or "borderline" performance were chosen based on general patterns from past seasons, not through an optimized or statistical process. This means the results could vary depending on how these cutoffs are adjusted. This calls for future test whether the same patterns hold across different conferences or seasons.

Finally, the absence of efficiency statistics like SP+ and EPA, known to enhance predictive accuracy [28, 29] could limit the depth of performance modeling.

5 Conclusions

This project was successful in proposing and quantifying what a "fraud" team is and creating a systematic way to identify them. Using performance metrics that highlight significant differences between perception and reality, the Fraud Score is capable of distinguishing contenders from frauds with high accuracy. Our framework provides a measure of how teams truly perform relative to historical championship-level benchmarks, offering insight into when hype is justified and when it is misleading.

Within our dataset, the evaluation of fraud score achieved consistent results across multiple seasons and aligned with expectations: teams with lower efficiency and narrower margins of victory tended to have higher fraud score, while dominant teams scored lower. The fraud scores not only separated contenders from frauds statistically but also held up under real-world validation, with case studies highlighting teams such as Mississippi State (2014), Texas A&M (2019), and Clemson (2022) as historically overvalued illustrate how perception and performance can diverge, supporting prior finds on ranking bias in the literature [1, 4, 5], while true contenders like Ohio State (2014), LSU (2019), and Michigan (2022) received low fraud scores.

From an ethical standpoint, the project also highlights the continuous disparities by biased ranking systems. As Edelman (2014) and Oplinger (2021) [23, 24] have argued, subjective rankings can reinforce disparities between Power Five and Group of Five programs, influencing recruitment, funding, and bowl access. Our findings of Fraud Score suggest that incorporating transparent, data-driven models could help minimize such biases. This aligns with broader calls in sports analytics for fairness, where algorithmic transparency ensures that evaluative models can be audited and improved over time rather than concealed by subjective judgment.

While these early results are promising, they should be interpreted with caution. The Fraud Score remains sensitive to how thresholds and perception metrics are defined, and its scope is limited by available efficiency statistics. Future research could refine the methodology by introducing other advanced efficiency measures like SP+, EPA, or Massey Ratings, and exploring whether similar approaches can predict postseason performance or playoff selection outcomes.

Overall, the project contributes to the growing body of work on data-driven fairness and evaluation in college football analytics. It demonstrates that while perception biases remain embedded in the ranking systems, other quantitative approaches can help reveal when team reputations align or fail to align with their on-field performance.

References

[1] B. Jay Coleman, Andres Gallo, Paul Mason, and Jeffery W. Steagall. Voter bias in the associated press college football poll. *Journal of Sports Economics 11(4):397-417*, 2010.

[2] E. Woodrow Eckard. Is the bowl championship series a cartel? some evidence. *Journal of Sports Economics 14(1):3-22*, 2011.

[3] Mark David Witte and McDonald Paul Mirabile. Not so fast, my friend: Biases in college football polls. *Journal of Sports Economics*, 11(4):443–455, 2010. doi: 10.1177/1527002510376617.

[4] Brent Hensley. Voter bias in the associated press college football poll. Honors thesis, University of California, Berkeley, 2015.

[5] Daniel F. Stone and Matthew Rod. Bias in the college football playoff selection process: If the devil is in the details, then so is the angel. *Marquette Sports Law Review*, 27(1):87–110, 2016.

[6] Andrew W. Nutting. Evidence from the top 25 ballots of ncaa football coaches. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=1964172, 2011.

[7] Wesley N. Colley. Colley's bias free college football ranking method: The colley matrix explained. *https://www.colleyrankings.com/*, 2002.

[8] Doug Drinen. Introducing srs. https://www.pro-football-reference.com/blog/index37a8.html, 2006.

[9] Sports-Reference. Srs calculation details. https://www.sports-reference.com/blog/2015/03/srs-calculation-details/, 2015.

[10] Arpad E. Elo. *The Rating of Chessplayers, Past and Present*. Arco Pub., New York, 1978. ISBN 0668047216.

[11] Nate Silver. Introducing nfl elo ratings. https://fivethirtyeight.com/features/introducing-nfl-elo-ratings/, 2014.

[12] Michael J. Mauboussin. *The Success Equation: Untangling Skill and Luck in Business, Sports, and Investing*. Harvard Business Review Press, 2012. ISBN 9781422184233.

[13] William N. Caballero and Katherine Zhao. An analysis of the ncaa college football playoff team selections using an elo ratings model. arXiv preprint arXiv:2403.03862, 2024.

[14] Daniel Berrar, Philippe Lopes, and Werner Dubitzky. Incorporating domain knowledge in machine learning for soccer outcome prediction. *Machine Learning*, 108(1):97–126, 2019. doi: 10.1007/s10994-018-5747-8.

[15] B. Deb, J. Fernández-Navarro, A.P. McCormack, D. Memmert, and R. Bornn. Finding repeatable progressive pass clusters and application in international football. *Journal of Sports Analytics*, 9(4):289–303, 2024. doi: 10.3233/JSA-220732.

[16] D. Fernandez, J. Fernandez-Navarro, J. Prats-Moya, and A. Fernandez-Ortega. Reporting of clustering techniques in sports sciences: a scoping review. *Electronic Journal of Applied Statistical Analysis*, 2024. doi: 10.12870/2067-4835.2024.3.29051.

[17] Deyu Wang and Hui Huang. Deep learning-based tennis match type clustering. *BMC Sports Science, Medicine and Rehabilitation*, 17(1), 2025. doi: 10.1186/s13102-025-01147-w.

[18] R. Pollard and G. Pollard. Home advantage in soccer: a review of its existence and causes. *International Journal of Soccer and Science Journal*, 3(1):28–38, 2005.

[19] Mark D. Risser, C. A. Calder, and Kiros Berhane. Home advantage in american college football games: a multilevel modelling approach. *Journal of Quantitative Analysis in Sports*, 7(3), 2011. doi: 10.2202/1559-0410.1344.

[20] T. Inan. The effect of crowd support on home-field advantage: Evidence from european football. *Annals of Applied Sport Science*, 8(3), 2020. doi: 10.29252/aassjournal.806.

[21] Christopher J. Boudreaux, Shane D. Sanders, and Bhavneet Walia. A natural experiment to determine the crowd effect upon home court advantage. *Journal of Sports Economics*, 18(7): 737–749, 2017. doi: 10.1177/1527002515595842.

[22] Mark W. Nichols. The impact of visiting team travel on game outcome and biases in nfl betting markets. *Journal of Sports Economics*, 15(1):78–96, 2014. doi: 10.1177/1527002512440580.

[23] Marc Edelman. The district court decision in o'bannon v. national collegiate athletic association: A small step forward for college-athlete rights, and a gateway for far grander change. *Washington and Lee Law Review*, 71(4):2319–2361, 2014.

[24] Oplinger. Bias in college football playoff selection process. Senior thesis, University of South Carolina, 2021.

[25] B. Radjewski. College football data api (v2). https://collegefootballdata.com/about, 2025. Accessed: 2025-08-29.

[26] Sport Reference. College football stats and history. https://www.sports-reference.com/cfb/, 2025. Accessed: 2025-09-13.

[27] College Football Playoff. College football playoff history and results. https://collegefootballplayoff.com/sports/2019/5/22/history, 2025. Accessed: 2025-10-08.

[28] Bill Connelly. Sp+ rankings after conference championship games. https://www.espn.com/college-football/story/_/id/28251219/sp+-rankings-conference-championship-games, 2019. Accessed: 2025-08-29.

[29] Alok Pattani. Expected points and epa explained. https://www.espn.com.au/nfl/story/_/id/8379024/nfl-explaining-expected-points-metric, 2012. Accessed: 2025-08-29.

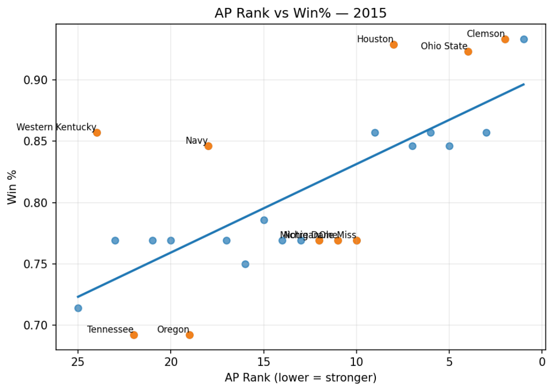
A Code Appendix (Github Reference)

The code repository of this project can be accessed here.

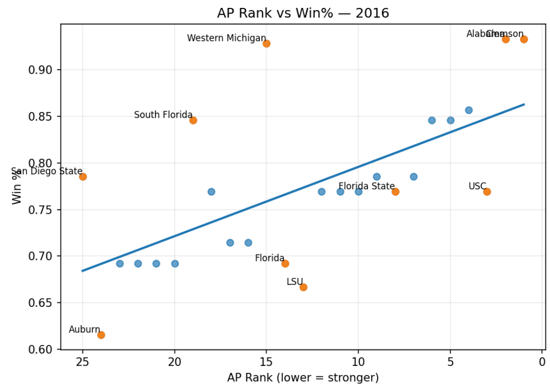
B List of All Teams from Each Cluster in Season 2019

- Cluster 0 teams (27): Ohio State, LSU, Alabama, Clemson, Wisconsin, Georgia, Penn State, Notre Dame, Oregon, Auburn, Oklahoma, Michigan, Florida, Utah, Iowa, Memphis, Navy, Washington, Texas, Baylor, Minnesota, UCF, Cincinnati, Air Force, Texas A&M, Boise State, Iowa State
- Cluster 1 teams (61): SMU, USC, Louisiana, Florida Atlantic, Oklahoma State, Kansas State, North Carolina, Virginia, Michigan State, Kentucky, Indiana, Tennessee, Mississippi State, Arizona State, Tulane, Washington State, California, Virginia Tech, Wake Forest, Louisville, Missouri, Nebraska, Ole Miss, Oregon State, Florida State, San Diego State, Buffalo, South Carolina, Wyoming, Temple, BYU, Tulsa, Western Kentucky, Duke, Colorado, Georgia Southern, Illinois, Boston College, Houston, Purdue, Marshall, UCLA, Texas Tech, Louisiana Tech, Western Michigan, Arkansas State, Utah State, Stanford, Ohio, Syracuse, West Virginia, Central Michigan, Miami (OH), Northwestern, Maryland, South Florida, Ball State, Army, Arizona, Fresno State, UAB
- Cluster 2 teams (30): Kent State, Troy, Liberty, Kansas, Arkansas, Georgia State, Colorado State, Charlotte, Coastal Carolina, Eastern Michigan, Florida International, Georgia Tech, Nevada, Toledo, North Texas, East Carolina, Northern Illinois, Vanderbilt, Rice, Rutgers, South Alabama, Texas State, New Mexico, UTSA, Bowling Green, New Mexico State, Old Dominion, UTEP, Akron, Massachusetts

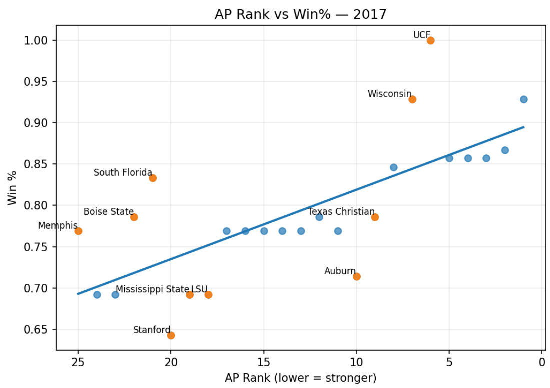
C AP Poll Ranking vs Win Percentage (2015-2024)



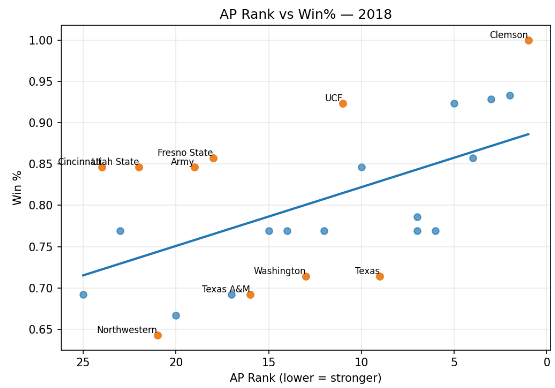
(a) 2015



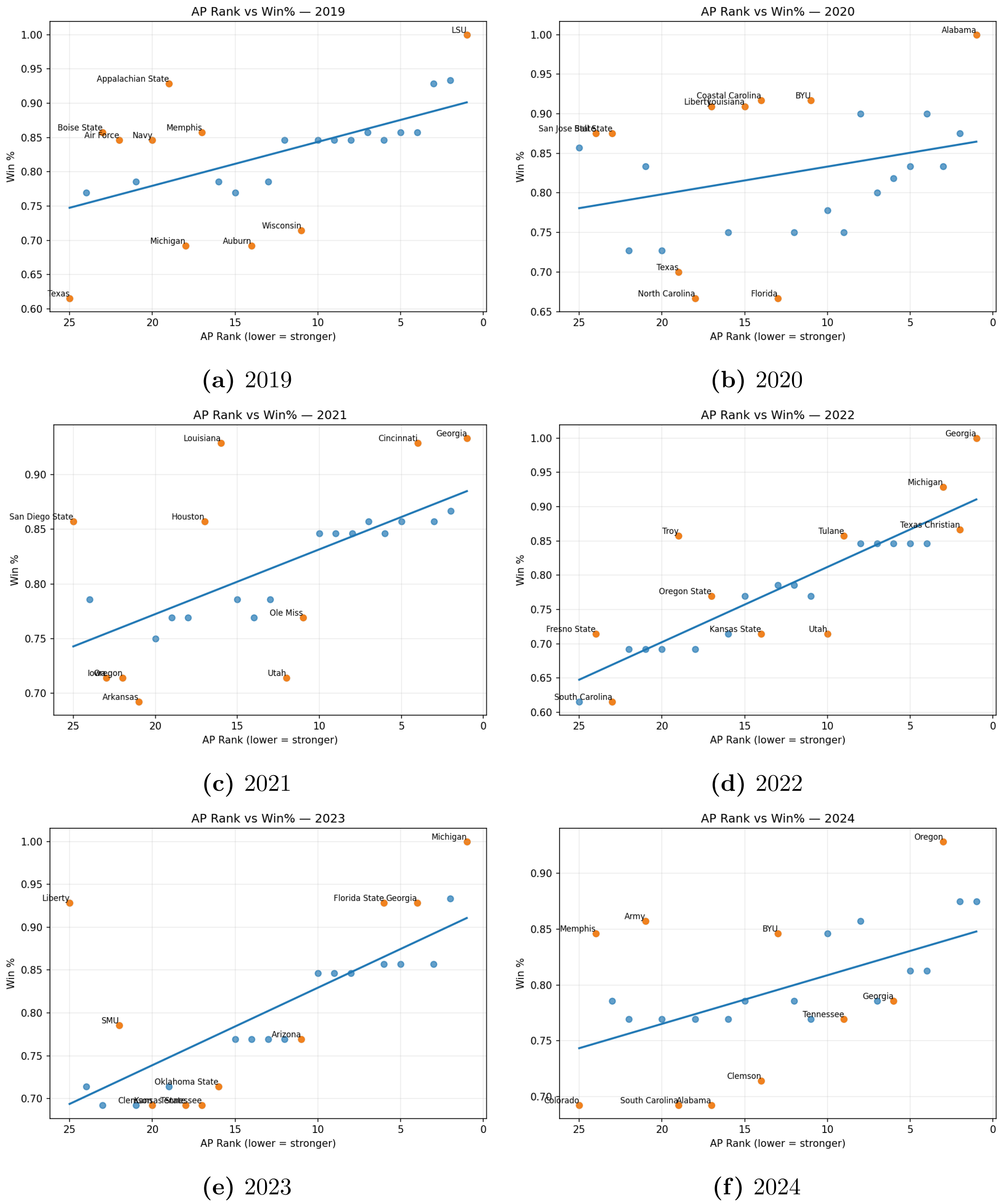
(b) 2016



(c) 2017



(d) 2018



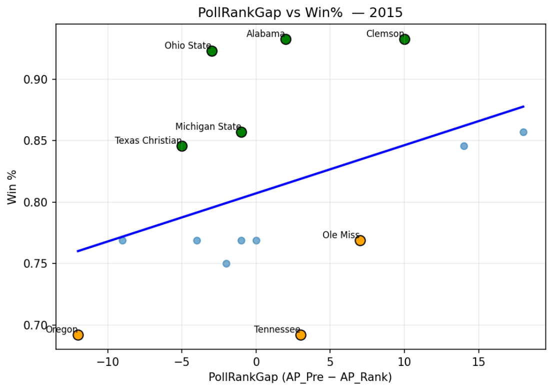
(a) 2019 **(b)** 2020

(c) 2021 **(d)** 2022

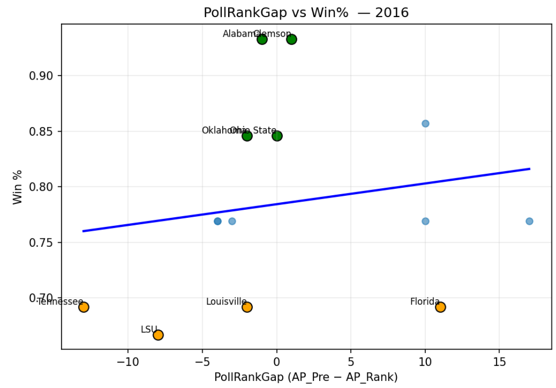
(e) 2023 **(f)** 2024

Figure 16: AP Rank vs Win% scatterplots for the 2015–2024 seasons

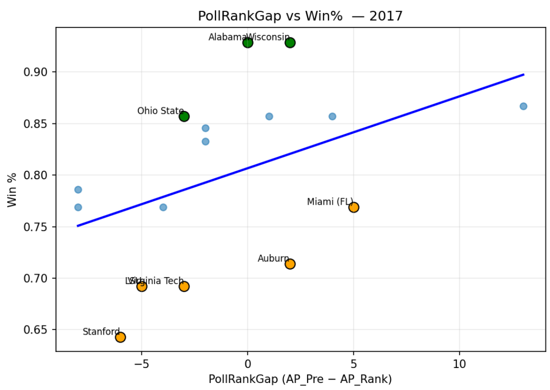
D PollRankGap vs Win Percentage (2015-2024)



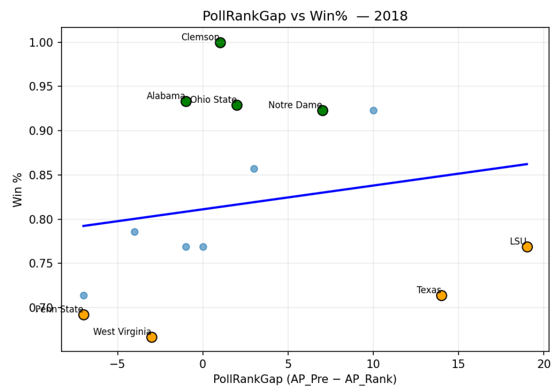
(a) 2015



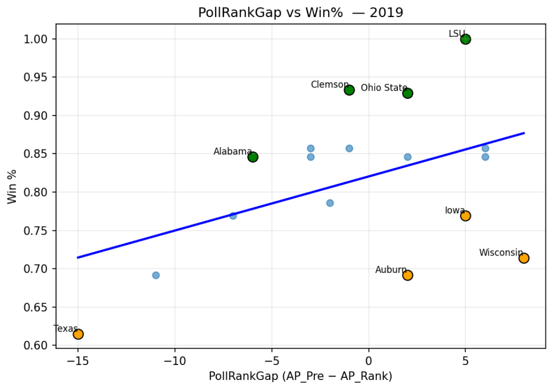
(b) 2016



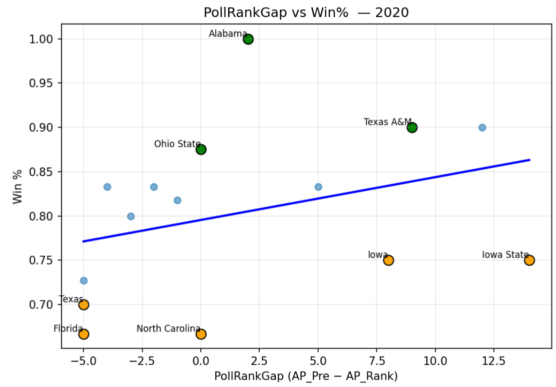
(c) 2017



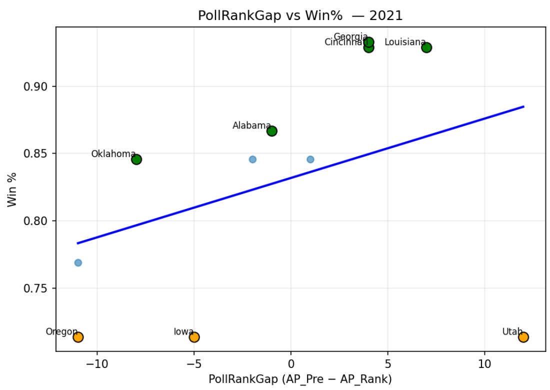
(d) 2018



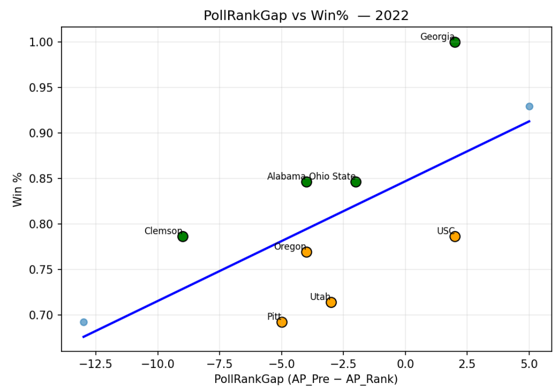
(a) 2019



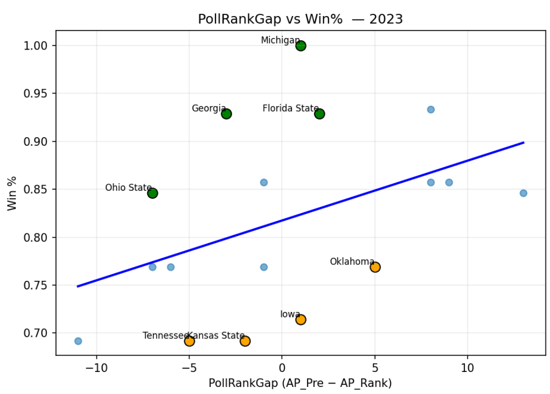
(b) 2020



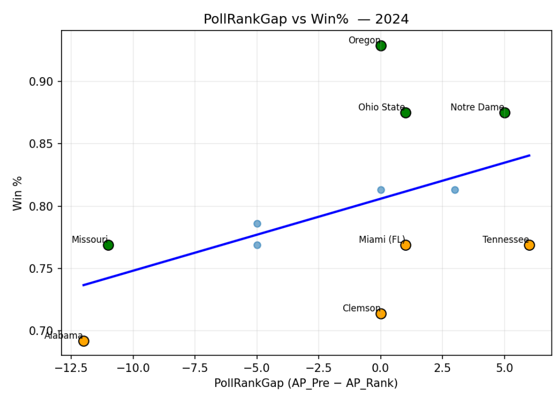
(c) 2021



(d) 2022



(e) 2023



(f) 2024

Figure 18: Poll Gap vs Win% scatterplots for the 2015–2024 seasons