Decision Dynamics in College Football Recruitment: An Analytical Approach to Predicting Commitment Patterns

A Thesis Presented to
the Faculty of the School of Engineering and Applied Science of
University of Virginia

in Partial Fulfillment of the Requirements for the Degree
Master of Science

By
Maryanna Lansing

Department of Systems Engineering
University of Virginia
May 2025

Approval Sheet

This Thesis
is Submitted in Partial Fulfillment of the Requirements for the Degree of
Master of Science

Author: Maryanna Lansing

This Thesis has Been Read and Approved by the Examining Committee:

Advisor: Matthew Burkett
Advisor: Robert Riggs
Committee Member: Matthew Bolton

Accepted for the School of Engineering and Applied Science:

Jennifer L. West, School of Engineering and Applied Science
May 2025

Dedication

I dedicate this thesis to my former mentor, advisor and professor: William T. Scherer. Without his guidance, support, and faith in me, the completion of this work would not have been possible. May he rest in peace.

Contents

List of Figures

List of Tables

1 Abstract

This research investigates the multifaceted decision-making processes behind high school football recruits' college commitment choices by integrating statistical hypothesis testing with advanced machine learning methods. Grounded in the growing importance of sports analytics in higher education and professional athletics, this study specifically addresses the gap in recruitment analytics for college football—a sport characterized by unique positional requirements, variable conference strengths, and diverse facility attributes. The work examines how intrinsic factors, such as a recruit's playing position and player tier, combined with extrinsic factors - like geographical distance from home, stadium capacity, and playing surface quality - collectively influence a recruit's commitment decision.

The study begins by establishing the context of sports analytics as a transformative field within the sports industry, where data-driven insights have revolutionized team management, performance analysis, and strategic planning. Although previous research has largely focused on in-game performance and financial aspects of professional sports, less attention has been given to the collegiate recruitment arena. This thesis aims to fill that void by developing a comprehensive model that predicts commitment outcomes across various demographic and contextual subgroups. Specifically, the research questions center on determining whether and how the distance from a recruit's home to a college, the capacity and quality of the college's stadium, the player's position, and their assigned tier interact to shape the final decision of where to commit.

To address these questions, a robust dataset was compiled comprising 132,522 data points from 10,734 unique recruits over a seven-year period (2017–2023). Data were aggregated from diverse sources, including recruiting databases, US Census statistics, collegiate performance ratings, and facility records, resulting in 71 distinct features. The methodological approach involved the Two-Sample T test 1to identify significant differences in key variables such as distance from home and stadium capacity between committed and non-committed groups.

Further analysis was conducted using Analysis of Variance (ANOVA) to examine group differences across categorical variables including playing position, conference affiliation, and player tier. The results of these tests revealed that recruits who ultimately commit to a college tend to come from homes that are, on average, closer to the institution compared to those who do not commit, with significant variations noted when data are stratified by position and player tier. Similarly, the quality of the college stadium—evaluated by its seating capacity and playing surface—emerged as a critical factor, particularly among higher-tier recruits. Interaction models were subsequently developed to explore the combined influence of these factors, revealing that the interplay between conference prestige, player tier, and positional demands contributes to nuanced patterns in commitment behavior.

On the modeling front, the thesis implements a series of machine learning approaches designed to enhance the predictive accuracy of recruitment outcomes. Multiple models, including Logistic Regression, Random Forests (both standard and tuned versions), Gradient Boosting, and a Two-Stage Decision Tree coupled with Random Forest, were evaluated. These models were trained on a pre-processed and standardized dataset with encoded categorical variables and imputed missing values to ensure data integrity. Cross-validation

techniques, specifically using Stratified K-Fold, were employed to assess model robustness. The primary performance metric used was accuracy, supplemented by log loss for models that output probability estimates. Among the various techniques, Logistic Regression consistently demonstrated high accuracy across most player positions, although specialized sub-models for certain positions, such as a Two-Stage model for positions with sparse data, were necessary to address imbalances.

The findings of this study underscore that both geographical and facility-related factors are significant predictors of recruitment decisions. Recruits from greater distances exhibit lower probabilities of commitment, a trend that is further moderated by their positional roles and perceived player tiers. In addition, the quality of a college's stadium, particularly its capacity and the type of playing surface, significantly influences a recruit's choice, with higher-tier athletes showing a pronounced preference for institutions with state-of-the-art facilities. These insights suggest that recruitment strategies can be optimized by tailoring outreach efforts and facility investments to the specific needs of different recruit segments. Moreover, the interaction effects observed between conference strength and player tier indicate that larger athletic programs may benefit from emphasizing both infrastructural advantages and strategic recruiting practices to attract top talent.

In conclusion, this thesis provides a novel contribution to the field of sports analytics and systems engineering by developing a predictive framework that elucidates the complex interdependencies influencing high school recruits' college commitment decisions. The integration of rigorous statistical testing with advanced machine learning techniques not only validates the significance of individual factors such as distance and stadium quality but also highlights the importance of their interactions with positional and tier-based distinctions. These findings have direct implications for collegiate athletic departments seeking to refine their recruitment strategies and for researchers aiming to further explore data-driven approaches in sports decision-making. Future research will extend this model by incorporating additional dimensions such as academic performance and long-term career outcomes, thereby offering a more holistic view of the factors that drive recruitment in college football.

2 Literary Review

2.1 Sports Analytics Overview

Sports are an important aspect of any culture. The sports industry is one of the largest in the world, generating an estimated revenue of $2.65 trillion [4]. An important part of the global sports industry is sports analytics. The sports analytics market was estimated at $3.78 billion in 2023, and is projected to grow to be between $16.45 billion and $32.31 billion by 2030–2032 [16].

But what is sports analytics? According to the research paper "Sports analytics: Designing a decision-support tool for game analysis using big data", "Sports analytics is the investigation and modeling of sports performance, implementing scientific techniques. More specifically, sports analytics refers to the management of structured historical data, the application of predictive analytic models that use these data, and the utilization of information systems, in order to inform decision makers and enable them to assist their organizations in gaining a competitive advantage on the field of play [3]." It is applying statistical methodologies, computational tools, and data sources to bigger concepts in sports.

Sports analytics has existed for many years, but it became a major part of the sports industry in 2002. During the 2002 Major League Baseball season, the Oakland Athletics revolutionized the game of baseball by using statistical analysis to determine the team's recruiting and drafting strategy. They were able to derive the key metric (OPS: On-base plus slugging) in determining player value and a team's winning ability [30, 31, 50]. Other teams in the league quickly caught on to the significance and impact that analytics could have on their team and season. From there, sports analytics really took off. Teams started developing models and metrics for all the different aspects of the sport, such as injury prevention, ticket sales, and contract analysis. Soon, other sports also caught on and started to apply similar methods. In basketball, they have studied a player's biomechanics to help understand different types of injuries and their impact on the player's development. In soccer, they have used video tracking data to determine which players have the greatest potential. In lacrosse, they have used location data to evaluate the likelihood a shot will be a goal.

Now sports analytics research isn't just occurring within each team but across leagues and within academics. Engineers, data scientists, statisticians, physicists, and more are using sports to explore new research topics and develop new methods and models. Table 1 describes the different research topics, data used and what category of sports analytics that are being studied across ten sports.

Table 1: Sports Analytics Research Across Different Sports

Sport	Approach	Data	Category	Reference
Baseball	Challenging conventional wisdom on batting order, which hitters will perform well in a scoring position and scoring rates of high-average hitters	Simulation data	Team Performance	[41]
	Predictive models of the odds a baseball player will get a base hit during any given game	Statistics, weather data, and ball park characteristics	Team Performance	[2]
	Tracking player movements to understand UCL injuries and work to reduce their risk	Video and radar data	Individual Performance	[19]
Basketball	Using Data Science and Machine Learning to study injuries and their impact on player management/development	Health data on basketball players from 2010 to 2020	Individual Performance	[46]
	Studying the affects of rule mining and injuries on player's salaries and team finances	Rules, injury data, sociodemographic data, and salary data	Team Finances	[47]
	Creating models to assist with team strategy and performance	Box score data and tracking data	Team Performance	[51]
Soccer	Effects of big data on team strategies and competitiveness	In game statistics, coaching data	Team Performance	[23]

	Predicting team performance for a season and approaches to detect an excellent central defender from a an average one	Tracking data and player statistics	Team/Individual Performance	[34]
	Use of data analytics in player recruitment	Video data, performance data	Recruiting	[18]
Ice Hockey	Use of data analytics to improve strength and conditions and see its effects on in-game player performance	Off-ice metrics, player statistics	Individual Performance	[27]
	Calculate metrics for passing lanes and player movement, passing effectiveness, and pressure	NHL tracking data	Team/Individual Performance	[40]
	Model and simulate analysis in talent identification	In game statistical player data	Recruiting	[25]
Volleyball	Using Markov chains and simulation to assist in in-game decision making	Coaching strategies, match statistics	Team Performance	[3]
Field Hockey	Modeling optimal practice schedules to increase team performance	Biometric player data, match results	Team/Individual Performance	[7]
Lacrosse	Evaluating the likelihood a shot will be a goal	Player tracking data, ball location data, passing data	Team Performance	[29]
Softball	Investigate trunk and pitching arm kinematics and their association with performance outcome	Biometric player data	Individual Performance	[17]

Track and Field	Explore the timing system that is more suitable for track and field competitions to achieve accurate ranging, reduce costs, and reduce errors	Timing data, location data, video data, imaging data	League Finances/ Individual Performance	[24]
Cricket	Developing solutions to track the ball's 3D trajectory	Sensor and radar data	Team/Individual performance	[22]

What all of these research topics point to is that there are two main categories in sports analytics: Performance and Business. Performance focuses on both the team and the individual athlete. Team and individual performance can be intertwined. When the individual does better, a team does better. However, when relating to team performance, sports analytics usually focuses on in-game decision making, expected wins, and overall strategy. Individual performance focuses on individual improvement, injury prevention, and optimal practice scheduling. Business focuses on recruiting, ticketing/sales, and team/league finances. Figure 1 shows graphically how these categories break down.

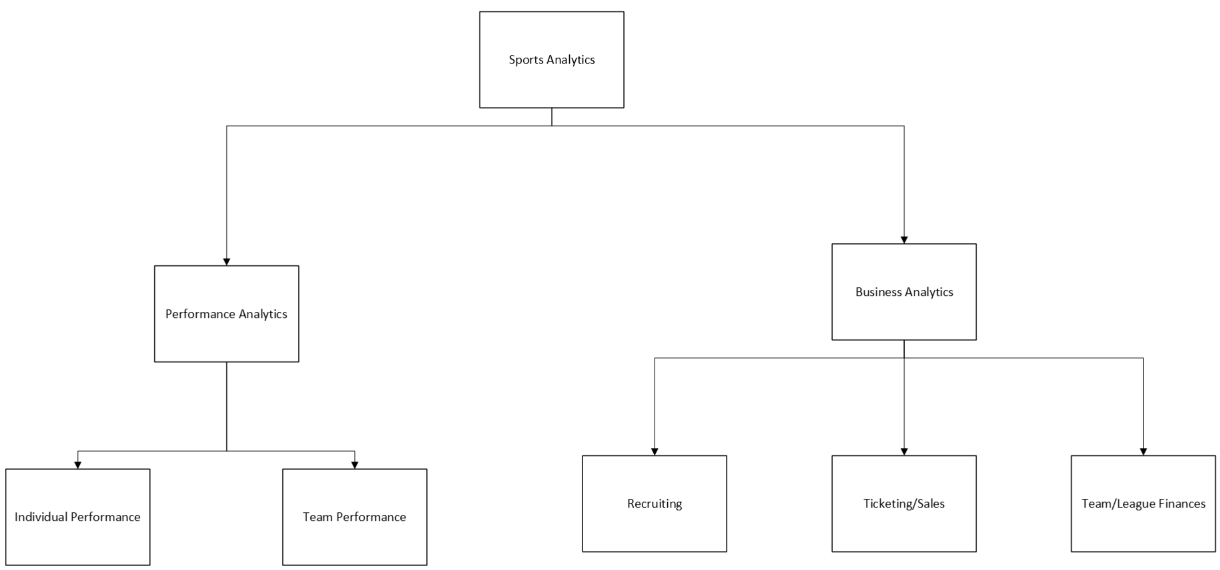

Figure 1: Sports Analytics Tree

2.2 American Football Analytics

Most sports are established in their analytics, while others are still developing what they can do with analytics. One sport that is still developing analytics is American Football. American football is among the largest sports in United States of America (USA). In 2023, the NFL (National Football League) brought in a revenue of $20.5 billion [5]. The NCAA (National Collegiate Athletic Association) had an estimated revenue of $1.3 billion in 2023, with nine programs making over $200 million [38]. In 2025, 128 million people watched the Superbowl and in the 2024-2025 NCAA football season 28 million people watch the College Football Playoff National Championship [9, 39].

American football is a complex game of teams within teams - an offense, a defense and special teams. The special teams exists within both the offense and defense teams. While the game is in play, 11 players from each team are on the field – offense vs defense. On offense there is the quarterback (QB), the running back (RB), the wide receiver (WR), the tight end (TE), and the linebackers (LB) (the center (C), the guard (G), and the tackles (T)). On defense there is the defensive tackle (DT), the defensive end (DE), the nose guard (NG), the linebackers (LB) (the middle linebacker(MLB), the weak-side linebacker (WLB), and the rusher (R)), the cornerback (CB) and the safety (S). Each of these positions have a very different and defined job. For example, the quarterback is the leader of the offense team. He calls the plays, passes the ball, and has to know where everyone is on the field. His mind has to be sharp and ready for anything. On the other hand, the running back has one job: to be as strong and fast as possible to run the ball from one end of the field to the other. However, all of the positions (offense or defense) work in coherence to either successfully run plays and, hopefully, score or read these plays and stop them.

With all the new types of data that are collected, teams and academics are diving into the analytics of American Football, trying to discover what metrics are significant to both recruiting and play, how they can make processes more efficient, and how to keep players safe. Table 2 describes different research topics, data used and what category of sports analytics that are being studied in the sport of American Football.

Table 2: Sports Analytics Research in Football

Approach	Data	Category	Reference
Predicting the Superbowl winner	Regular season data	Team Performance	[43]
Predicting the true value of a player	Play by play data, contracts, snap counts and draft combines	Individual Performance/ Team Finances	[21]
Modeling expected points and providing insights for in game decision making	Player tracking data, in game statistics	Team Performance	[36]
Modeling a scale rating systems for offensive players	Game data	Individual Performance	[48]
Applying a plus-minus valuation system	Game data, tracking data, team records	Individual/ Team Performance	[44]
Using data analytics to make play calling more efficient and effective	in Game data, team record, player data	Team performance	[15]

Most analytics have been done for in-season and in-game evaluations. An untapped area of research in American football is recruitment. However, recruiting is essential for team success and should be explored in research. As discussed earlier, baseball's great analytical achievement was finding the correct metric to draft the best players. American Football has yet to do this. There are many notable examples of NFL missing the mark but two specific ones are Tom Brady and Brock Purdy. Tom Brady was the 199th overall pick (a six round selection) in the 2000 Draft [52]. He would go on to become arguably the greatest QB of all time, winning seven Superbowls. Brock Purdy was the 262nd overall pick in the 2022 draft, earning him the nickname "Mr. Irrelevant" [8]. Within 2 seasons, he led the San Francisco 49ers to the Superbowl in 2024. On the other hand, there have been many busts on early draft picks and many booms on early draft picks. It is still a guessing game. College football is no different but has the added problem that the teams aren't drafting the players. Instead, they recruit players who can say no and change their mind. Also, college football teams aren't competing for an order where they may get first chance at player, but rather, they compete against multiple teams all going for the same player at the same time.

This paper explores the question: "What effects a high school recruit's decision when deciding on a school?" Back in 2018-2019, the University of Virginia (UVA) Systems Engineering Department had two capstone projects who worked with the football team to develop recruiting and performance models. The recruiting models focused on predicting the likelihood of specific players coming to UVA. As an add on, the recruitment capstone team created a model to find "Diamonds in the Rough": two and three star players who have the grit and potential to play above expectations [15, 42]. The performance models focused on helping with in-game decision making. They helped the team decide whether Virginia should go for it on fourth down and analyzed data from sensors that players wore beneath their pads to improve individual performance [15, 42].

This paper focuses on creating an updated model that predicts the likelihood a recruit attending a certain college. It is not specific to the University of Virginia but predicts against all conferences, schools, and recruits. Also, the model will not include any financial data, which is difficult to find as it doesn't have to be disclosed by schools. This model is an overall model of the data. Models for specific positions and tiers of players are also developed to better predict commitment in a complex sport. This paper explores the significance of certain features impacting a recruit's choice. These features include the position the recruit plays, the distance of the school from a recruit's home, the tier of a player (top, middle, and bottom), the conference of the school, and the quality of a school's stadium.

3 Research Questions and Justification

College football continues to be a very niche and unexplored area in sports analytics. Most researchers focus on pro sports, letting that research trickle down to college sports. However college athletics face different challenges, scenarios and rules. These differences make for interesting and unique research.

As stated in the last section, this paper explores the question: "What effects a high school recruit's decision when deciding on a school?" This was explored through statistical testing and statistical modeling.

During the first semester of the 2024-2025 academic year, the Systems Engineering department, with the help of the consulting company TapHere!, created a team to work with UVA football, exploring different recruiting analytics for them. One job analyzed the likelihood of a recruit to commit to a certain school, understanding what metrics were significant for the recruits' choice. The team scrapped all the data that was used in the analyses across this paper from online. Multiple sources were used including 247Sports, US Census data, College Football Api, Massey Ratings, US News and World Report and Catapult. Seventy-one different features were collected. Table 3 shows all the 71 features and their names in the dataset.

Table 3: Features in Full Dataset

Type of Feature	**Name of Features**
Continuous Features	247Sports ID, row_count, Overall Ranking, Height, Weight, Composite Score, Position Rank, rank_state, county_fips, home_lat, home_lng, population, density, zips, TOTAL_POP, TOTAL_MALE, TOTAL_WHITE, TOTAL_NATIVEAMERICAN, TOTAL_ASIAN, TOTAL_PACIFICISLANDER, TOTAL_BIRACIAL, TOTAL_HISPANIC, Unemployment_rate_2022, Median_Household_Income_2021, Median_HH_Income_Percent_of_State_Total_2021, location.zip, college_lat, college_long, location.elevation, location.capacity, location.year_constructed, Record %, SoS, academic_rank, dist_from_home
Categorical Features	First Name, Last Name, Position, High School, City, initials_state, Commitment Status, Recruiting Class, official_visits, unofficial_visits, school_camps, tested_offer, coach_visits, committed_to, off_match, unoff_match, camp_match, name_state, county_name, ranking, RUCC_2023, abbreviation, conference, classification, location.name, location.city, location.state, location.timezone, location.grass, location.dome, National, Liberal, Regional, in_state, player_tier

Please look at Table 26 in the appendix for a definition of each of these features. For categorical features, there is also a definition for each category. Overall, 132,522 data points were collected. This data is comprised of recruits from 2017 - 2023 with 10,734 unique recruits in the datasets. Below is a table showing examples of the data points.

Table 4: Example of Dataset

Overall Ranking	Position	Height	Weight	Composite Score	...	in_state	player_tier
2722	RB	70	170	0.796	...	0	Bottom Tier
2722	RB	70	170	0.796	...	1	Bottom Tier
1855	DT	73	270	0.817	...	1	Mid Tier
1855	DT	73	270	0.817	...	1	Mid Tier
795	SDE	76	260	0.861	...	0	Mid Tier
795	SDE	76	260	0.861	...	0	Mid Tier

Because "What effects a high school recruit's decision when deciding on a school?" is a very multifaceted question, to better understand that question, and dive more into the specifics of high school recruits' commitment choices, this question was broken down into two smaller research questions:

1. Does the position of a player and distance from the player's home effect where a player commits to? Are these effects influenced by the conference? Are these effects influenced by the tier of the player?
2. Does the quality of the stadium of a college influence a players choice? Does the position of the player effect this? Does it depend on the tier of the player?

Each question focuses on the significance of certain features impacting the recruit's choice. The variable that was predicted and tested against is committed_to. The features include the position the recruit plays, the distance from a recruit's home a school is, the tier of a player (top, middle, and bottom), the conference of the school, the location of a team's stadium, the stadium's capacity, and the stadium's grass or turf field.

To further understand the features that predict a high school recruits' commitment choices, multiple models were created. One model was an overall model that predicts the variable committed_to with no grouping. It took in all the data and predicted across all schools, conferences, positions, etc. The next models also predicted the variable committed_to but grouped the data by position and player tier. These groups of data were used to create specific position models. Position based models were explored because American football is a very position oriented sport. The position of a recruit has a lot of effect on what choices the recruit has, what their height and weight are going to be in the data, and there overall composite score.

The final models predicted the variable committed_to and used data grouped by player_tier. This model showed the differences between different tiers of players in where they can commit and what effects their decisions. The caliber of the player can affect what offers a player will get, what schools they can look at, how many offers they receive and what conferences they focus on.

Figure 2 and Figure 3 show the total data points for each category in the main variable groups explored in the research questions and the models (player_tier and Position). The majority of the data points are categorized by mid tier recruits (see appendix table 28). Since there are so many position options, the data is pretty well spread out among the positions. However, there are three positions that have very few data points: FB has only 20, LS has only 18, P has only 37. This was taken into account when understanding their models.

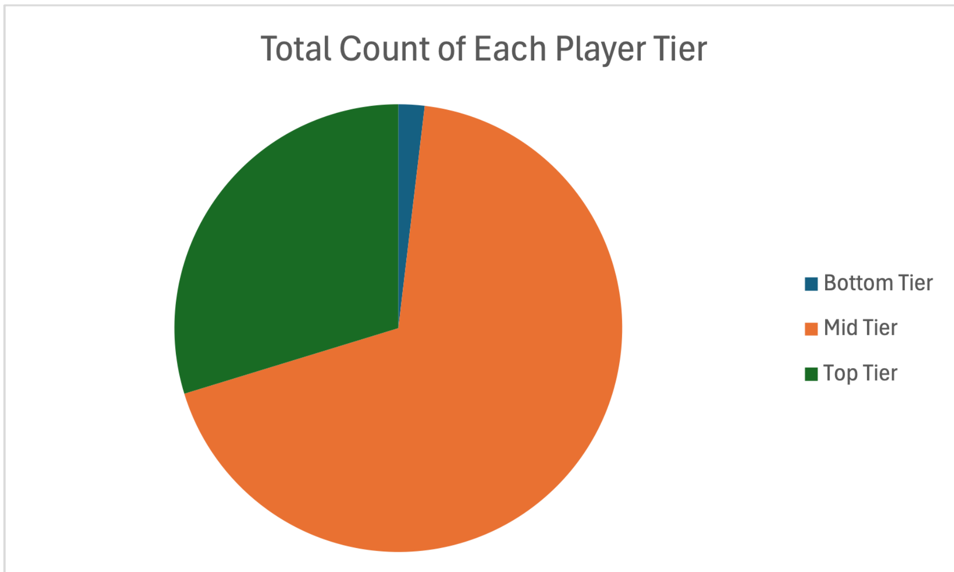

Figure 2: Count of Player Tier

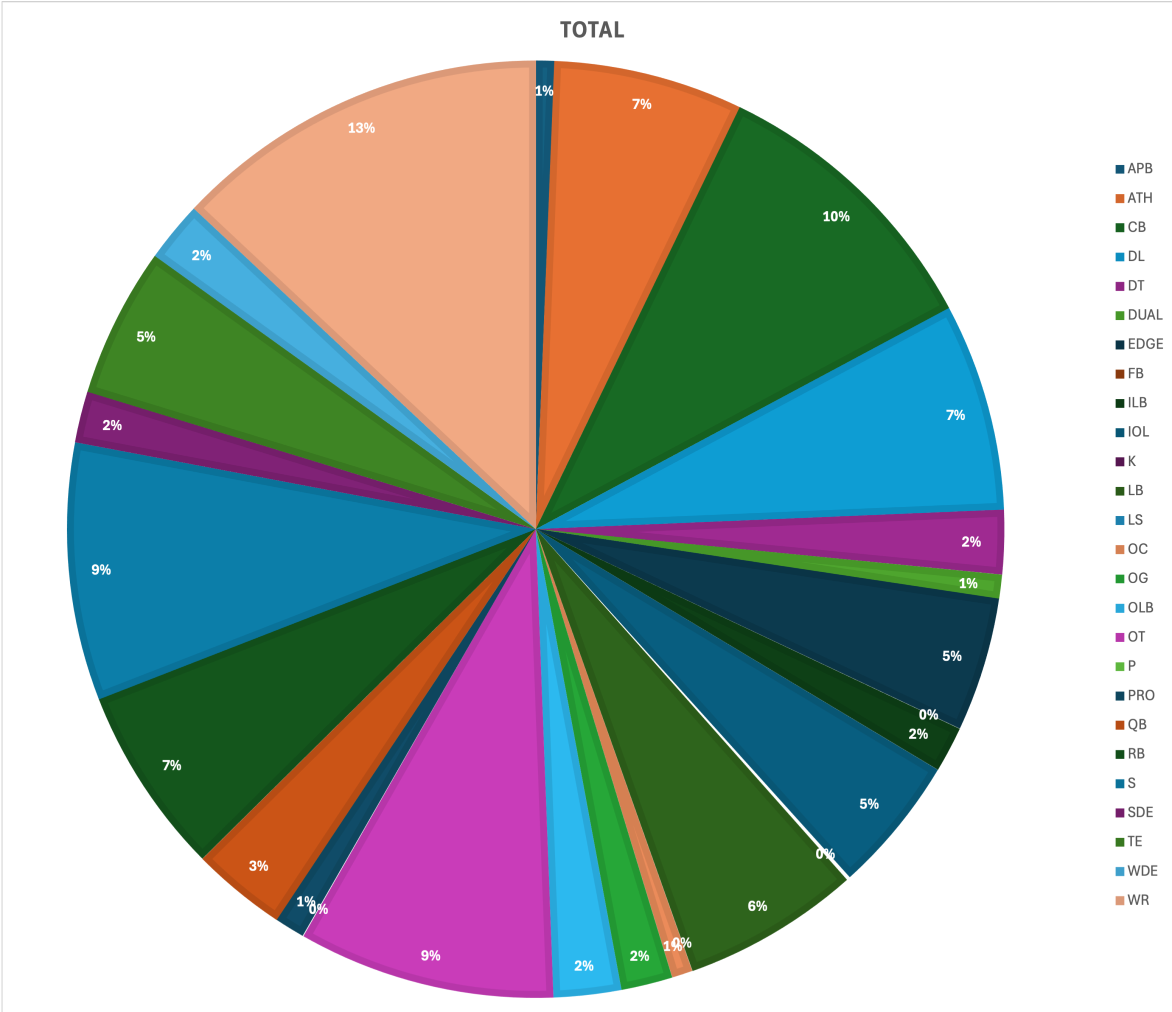

Figure 3: Count of Position

Figure 4 shows the breakdown of committed_to by conference. Overall 8.38% of the data is commitment data points to schools. The other 91.62% is all the schools that recruits looked at but decided not to go to. Since the data is so skewed in its majority and minority classes, there is likely more to learn from the data about why a recruit didn't go to a school then there is from why they did go to a school.

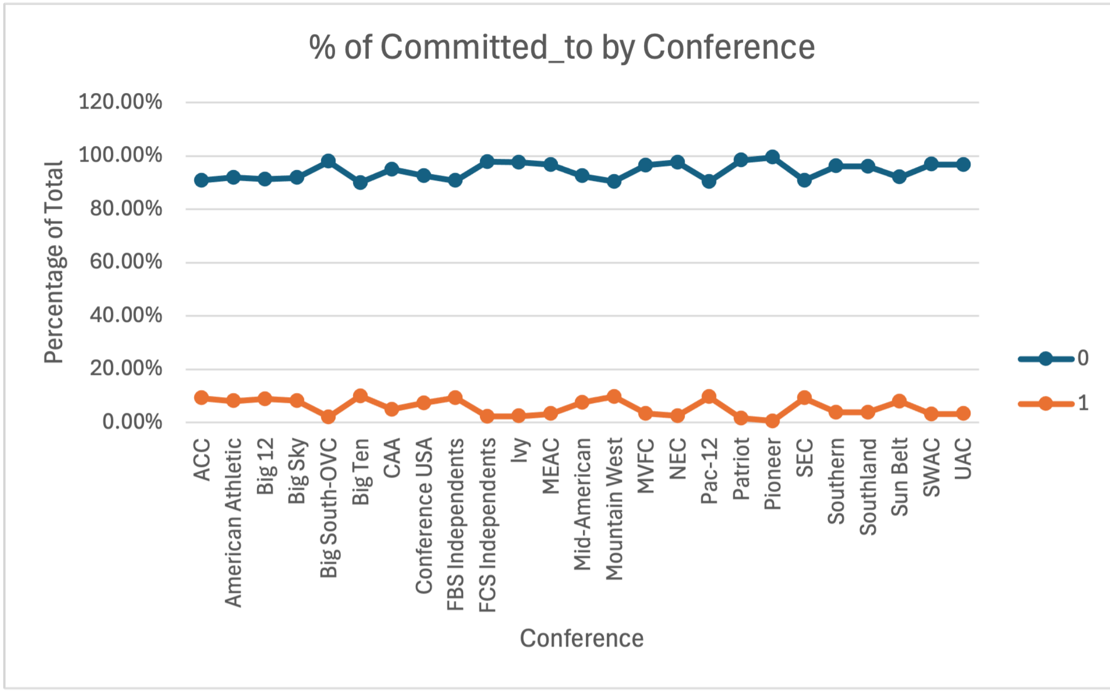

Figure 4: Committed_to by Conference

3.1 Question 1: Does the position of a player and distance from the player's home effect where a player commits to? Are these effects influenced by the conference? Are these effects influenced by the tier of the player?

These questions explore the effect of distance on a recruits decision. Table 5 shows the average distance from home for both commitment status: true (committed) or false (did not commit). The difference in the averages is 147.002 miles. This table and difference covers all recruits regardless of any grouping. This is a large difference but what is more interesting is that the average distance from home for school's recruits actually committed to is 461.158 miles. This may seem like a great distance but in reality, it shows that recruits are staying in their region of the country if not their state. Furthermore with the consolidation of conferences, teams may be thousands of miles apart.

Table 5: Average Distance from Home by Commitment

Committed_to	Average of dist_from_home
0 (False)	608.161
1 (True)	461.158

Because many factors go into this overall value, it is prudent to look deeper into certain groups. Table 6 shows the average distance from home for both commitment status being true or false for each position. This data was used to test to see if distance from home is significant and just how significant. The grouping was broken down even further by looking at each position within in each conference and by tier of player.

Table 6: Average Distance from Home by Position and Commitment

Position	**Avg(dist) for committed_to = 1**	**Avg(dist) for committed_to = 0**
APB	565.829	649.142
ATH	451.230	556.915
CB	503.874	650.0251
DL	439.867	570.016
DT	405.035	558.792
DUAL	546.218	691.110
EDGE	476.251	610.195
FB	826.801	618.670
ILB	463.711	601.793
IOL	429.849	581.449
K	574.890	762.506
LB	485.237	611.828
LS	573.462	875.463
OC	370.755	573.709
OG	388.921	582.899
OLB	480.912	641.639
OT	395.314	587.143
P	811.781	747.491
PRO	571.912	709.293
QB	593.2710	652.242
RB	510.241	596.268
S	444.368	611.804
SDE	416.006	577.696
TE	423.294	600.509
WDE	452.861	654.440
WR	454.167	627.337

3.2 Question 2: Does the quality of the stadium of a college influence a players choice? Does the position of the player effect this? Does it depend on the tier of the player?

This question explores the effect of the quality of a stadium on a recruits' decision. Table 7 shows the average stadium capacity for both commitment status being true or false. The difference in the averages is 4847.589 seats. This covers all recruits and stadiums regardless of any grouping. The largest stadium in NCAA football is the University of Michigan's stadium. It has a capacity of 107,601 seats. Sam Houston has the smallest stadium with a capacity of 14,000 seats [26].

Table 7: Average of Stadium Capacity by Commitment

Committed_to	**Average of location.capacity**
0 (False)	51546.875
1 (True)	56394.464

Again, it is prudent to look deeper into certain groups, to understand the effects of the quality of a stadium. Table 8 and Table 9 show the average stadium capacity for both commitment status being true or false for each position and for each player tier. This data was tested to see if stadium capacity is significant and just how significant.

Table 8: Average of Stadium Capacity by Position and Commitment

Position	Avg(capacity) for committed_to = 1	Avg(capacity) for committed_to = 0
APB	58616.812	65850.030
ATH	47262.679	52145.148
CB	53883.628	58798.792
DL	50966.248	55916.059
DT	58208.003	65630.465
DUAL	53868.282	59852.972
EDGE	52098.394	54913.277
FB	39356.438	42556.250
ILB	55770.447	64978.922
IOL	47362.351	54023.764
K	42189.343	66163.170
LB	49119.059	50304.450
LS	61844.462	58648.000
OC	51273.375	62370.082
OG	54925.483	59780.013
OLB	53398.724	59888.710
OT	50539.655	57731.321
P	48064.615	71963.545
PRO	48843.127	55221.740
QB	46863.576	49346.562
RB	51258.008	54136.0815
S	51887.992	55848.369
SDE	55678.149	62774.392
TE	50963.045	57387.335
WDE	60281.449	61150.866
WR	52264.557	54583.747

Table 9: Average of Location Capacity by Tier and Commitment

Tier of Player	Avg(capacity) for committed_to = 1	Avg(capacity) for committed_to = 0
Bottom Tier	24720.824	29162
Mid Tier	45112.759	51638
Top Tier	67077.774	84066

Whether or not a stadium has a grass or turf field was also used to evaluate the quality of a stadium and its effects on a recruits' decision. Having a grass field is important because it improves ball control and research has shown that less injuries and less severe injuries occur on grass fields than on artificial turf fields [28]. Figure 5 shows the percent of commitments to schools with a grass field and the percent of commitments to a school with a turf field. Grass fields are represented by the orange and turf fields are represented by the blue.

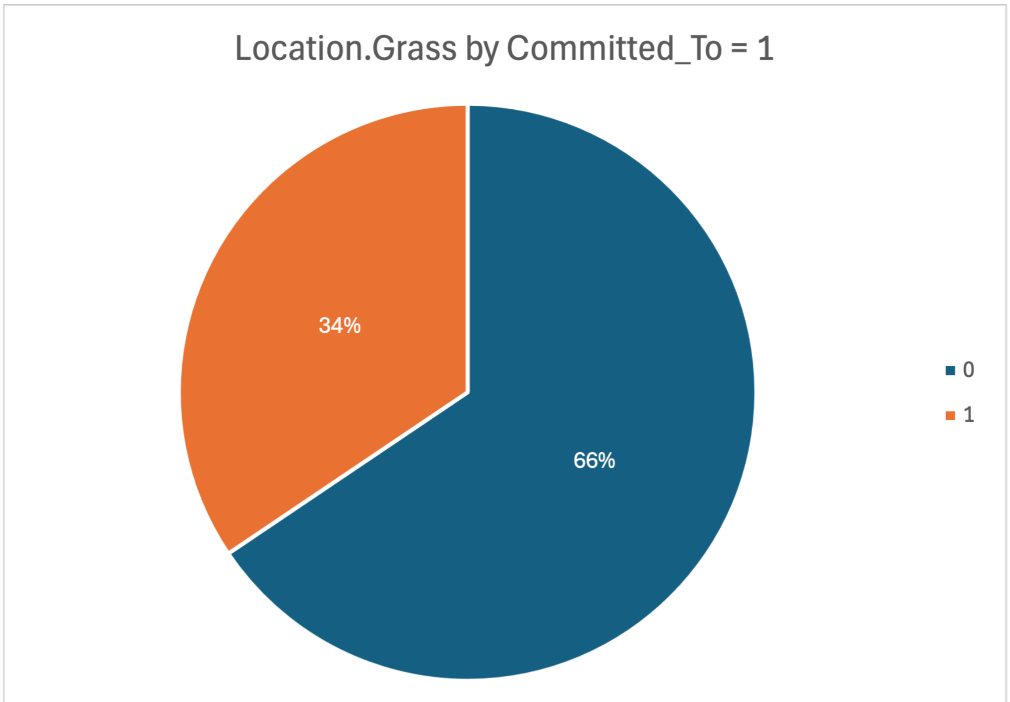

Figure 5: Percent of Schools Recruits Committed to with Grass Fields or Turf Fields

4 Methodology

This section describes how the research questions were answered, what types of models were run, and why certain model and the validation methods were picked in deciding on the best model for predicting committed_to overall, grouped by position and grouped by player_tier.

To answer the research questions, statistical hypothesis tests were ran. The tests started broad with overarching significance and then got more and more specific.

4.1 Research Questions

4.1.1 Question 1: Does the position of a player and distance from the player's home effect where a player commits to? Are these effects influenced by the conference? Are these effects influenced by the tier of the player?

For these questions the features focused on were dist_from_home, committed_to, position, conference and player_tier. Why focus on these features? Over 50% of students choose to attend a school within 100 miles of home [1, 49]. The goal was to see if the high school football recruits had a very specific reason to look at a school held true to this as well. Also because a recruit is looking to play football and attend school, they have more factors going into there decision. To see if this desire to be closer to home was affected by conference, position of the player or the caliber of the player, it was important to consider conference size, strength of their sports, and money they had. This paper does not focus on the financial side of things but it is good to point out the conferences with more money can recruit more players and better players. With that being said, are these bigger and better conference still getting players who are closer to home or are they recruiting from farther away? In addition, each position in football has a very different type of player and different type of role. For example, a team wants a 230 lb, 6 foot tall guy for a linebacker, but for a kicker if he is that heavy, he probably isn't very good. Also different years, teams are looking for different positions. Therefore a recruit might be affected by the needs of the team for that year. Finally, if the recruit is a higher caliber of player, he has more choices and bigger schools looking at him. This could affect if he chooses to stay closer to home or attend the best school he is being recruited to even though, it is far away.

To see what effects all these features may have, multiple statistical tests were run. First, the significance between dist_from_home and committed_to was tested. A Two-Sample T Test was used. The hypothesis tested was the mean of distances from home for players who committed is different from those who did not commit. An ANOVA test was then ran for position, conference and player_tier with committed_to against dist_from_home. The hypothesis tested whether the distance from home is influenced by player position, conference, and player tier. Two multi-factor ANOVA tests, one where the data is group by position within each conference and one where the data is group by position within each player_tier, were run. The hypotheses tested with these tests were interaction effects exist between commitment status and each of the factors (position, conference, player tier) and the main effects of commitment status, conference, player tier, and position significantly predict the distance from home. The final test was an overall multi-factor ANOVA test with interactions to see if there was significance between position, player_tier and conference when using dist_from_home to understand committed_to. Through all of these tests, it was shown that the distance from home matters to recruits and that this is different by position, player_tier and conference.

4.1.2 Question 2: Does the quality of the stadium of a college influence a players choice? Does the position of the player effect this? Does it depend on the tier of the player?

For these questions, the features that were focused on were location.capacity, location.grass, committed_to, position, and player_tier. So why focus on these features? Even though it doesn't seem like it, since a recruit can play anywhere, the stadium in which a team plays is very important. If the field is grass or not can be determinant on possible injuries. Studies have shown that players are more likely to be injured playing on turf [28]. Also, home field advantage has been proven to be real. In American football (NFL specifically), on average, teams win 57.6% of their home games [37]. A study done at University of Bristol on the 2008-2009 NCAA Football season showed that 73% of the teams had a higher winning percentage at home than away [53]. One of the factors used in this study to model the home field advantage was attendance or the crowd size for each game. Stadium size affects how many people can attend.

Therefore, the objective is to explore the influence of the quality of a stadium on player commitment decisions across different player positions and tiers. To meet this objective, multiple Two-Sample T tests and Chi-Square Tests of Independence were run. First, the significance between location.capacity and committed_to was tested using a Two-Sample T test. This specific test was used because it evaluates if there is a difference in the means. Then, the significance between location.grass and committed_to was tested. A Chi-Square Test of Independence was used because both location.grass and committed_to are categorical, binary variables. Location.grass = 1 is it is true a stadium has a grass field. Location.grass = 0 is it is false that a stadium has a grass field. Finally tests where the significant location features were group it first by position, and then second by player_tier were tested. Through all of these tests, it was shown that recruits prefer bigger stadiums with grass field, and that this preference is different by position and player_tier.

4.2 Models

Each model was coded in python. The code read in the data, cleaned it up and made any transformations necessary. The cleaning process included:

- Missing values in the target variable 'committed_to' were dropped to ensure data integrity for model training.
- Columns with problematic names were renamed for consistency and ease of access.
- Rows with specific unwanted entries in 'Record_Percent' and 'SoS' columns were removed. This was done to make them both fully numeric columns. This allowed them to be used as a continuous variables in the model instead of categorical variables.
- Conversion of certain columns to a numeric data type to ensure they were suitable for

statistical analysis.

- Dropped columns that are entirely NaN and those identified as overly specific or irrelevant.
- Correlated columns were also removed to prevent multicollinearity.

The transformations included:

- Categorical data encoding: The target variable 'committed_to' was transformed from a categorical to a numerical format using LabelEncoder.
- Imputation: Missing values in numerical and categorical features were filled using the mean and the most frequent value respectively.
- Data scaling: Standardization was applied to the feature set to normalize the data.

4.2.1 Overall

The Overall Model used machine learning techniques tailored to predictive analytics in sports. By employing a variety of models and hyperparameter tuning, the best possible model was identified based on accuracy and log loss metrics. A detailed evaluation and a use of cross-validation contributed to the reliability and validity of the model results, making this approach highly valuable for making informed decisions in player recruitment and team building.

Feature selection occurred inherently in some of the models used (e.g., models that utilize feature importance like Random Forests and Gradient Boosting) so it was not explicitly conducted.

To perform cross validation, the python method StratifiedKFold was used. It split the data, which ensured that each fold was a good representative of the whole by maintaining approximately the same percentage of samples of each target class.

Multiple models were considered, including Logistic Regression, Random Forest, XGBoost, SVM, and Gradient Boosting. Each model was set up with specific hyperparameters that were tuned using the python method RandomizedSearchCV. This method of tuning introduces randomness in the selection of parameters, which can help in discovering the best model configurations more efficiently than grid search.

Each model was trained and evaluated using the cross-validation setup. The RandomizedSearchCV was particularly useful here as it not only helped in tuning the parameters but also performed the training across different folds automatically, thus providing a robust estimate of the model's performance.

Accuracy of the model will be the primary metric to assess model performance, which provides a straightforward measure of how often the model predicts correctly. Accuracy is used instead of other indicators like AIC and BIC because to develop these models, machine-learning methods are being used.

Log loss is also calculated for models capable of producing probability estimates (like Logistic Regression and Gradient Boosting), providing a measure of uncertainty or confidence in the predictions. It also was used to evaluate the best model in cases of very similar or equal accuracy.

4.2.2 By Position and By Player Tier

Both the models for Position and Player Tier were tailored to handle specific subsets of data (positions and tiers), employing robust statistical techniques to ensure that the models were both accurate and relevant to their respective groups. The use of advanced ensemble techniques and the strategic two-stage modeling approach demonstrated a sophisticated handling of the prediction tasks, potentially suited to the complexities and nuances of predicting player commitments in college football.

These methods provided a good balance between statistical rigor (through feature selection and ensemble modeling) and practical applicability (through accuracy metrics and detailed model evaluations). This dual focus is essential in sports analytics, where both the statistical significance and actionable insights are crucial for making informed decisions.

The model used ANOVA F-test (f_classif) within the the python method SelectKBest method to identify the top 15 features. ANOVA F-test is used to find features that have a strong relationship with the target variable by checking if the means across multiple groups differ significantly, suitable for this categorical target variable.

All of the following models were trained and tested to evaluate the best method for each position:

- Logistic Regression: This model is used for its simplicity and effectiveness in binary classification tasks. It estimates probabilities using a logistic function, which is particularly useful for understanding the impact of each feature on the likelihood of outcomes.
- Random Forest: This ensemble method uses multiple decision trees to improve classification accuracy and control over-fitting. It is good for handling large datasets with higher dimensionality.
- Tuned Random Forest: Adjustments include setting class weight to 'balanced' and limiting the tree depth, which helps in addressing class imbalance and preventing overfitting.
- Gradient Boosting: Another ensemble technique that builds trees sequentially, with each new tree attempting to correct errors made by the previous ones. It's often praised for its predictive accuracy.
- Two-Stage Decision Tree and Random Forest Model: Begins with a Decision Tree to filter the data, followed by a Random Forest model trained on the subset of data selected by the Decision Tree. This staged approach can refine the focus on harder-to-classify instances, potentially improving model performance on complex or imbalanced datasets.

Gradient boosting was particularly noteworthy for handling potentially complex relationships within higher or lower-tier players, which might involve subtler distinctions in player characteristics influencing their commitment decisions.

Accuracy is used to evaluate each model. This metric is straightforward but does not consider the class distribution, which can be problematic in imbalanced datasets. Accuracy is used, instead of other indicators like AIC and BIC, because to develop the subject models, machine-learning methods were used. The F-score was also calculated for all the models. It was used as another way to evaluate the best model in cases of very similar or equal accuracy.

5 Results and Analysis

5.1 Research Questions

5.1.1 Question 1: Does the position of a player and the distance from the player's home effect where a player commits to? Does it depend on the conference? Does it depend on the tier of the player?

To answer Research Question 1, the following tests were run: Two-Sample T-Test for mean comparisons, and ANOVA for group mean differences, including interaction effects.

To run the Two-Sample T-Test, the dist_from_home data was group by committed_to=0 and committed_to=1. Running of the test showed there were significant differences in distances, indicating distance impacts commitment decisions. Table 10 shows results of the Two-Sample T-Test.

Table 10: Two Sample T-Test

T-Stat	-28.746
p-value	0.000

Since distance impacts commitment decisions, the nuances of that impact needed to be evaluated. To understand these nuances, the data was divided into three more groupings: Position, conference, and player_tier. To understand the significance of these groupings, three separate ANOVAs were run. Table 11, 12, and 13 show the results for the ANOVA test for Position, conference, and player_tier.

Table 11: ANOVA Results for Position

	sum_sq	**df**	**F**	**PR(>F)**
C(committed_to)	2.202e+08	1	773.551	9.388e-170
C(Position)	1.372e+08	25	19.283	5.905e-86
C(committed_to):C(Position)	1.317e+07	25	1.850	5.991e-03
Residual	3.770e+10	132469	NaN	NaN

Table 12: ANOVA Results for Conference

	sum_sq	df	F	PR(>F)
C(committed_to)	2.674e+08	1	1084.747	6.238e-237
C(conference)	5.130e+09	24	867.0900	0.000e+00
C(committed_to):C(conference)	6.936e+07	24	11.724300	1.043e-45
Residual	3.265e+10	132471	NaN	NaN

Table 13: ANOVA Results for Player Tier

	sum_sq	df	F	PR(>F)
C(committed_to)	1.873e+08	1	659.142	5.210e-145
C(player_tier)	1.837e+08	2	323.320	8.421e-141
C(committed_to):C(player_tier)	1.721e+07	2	30.286	7.075e-14
Residual	3.765e+10	132515	NaN	NaN

Significant effects were found for Position, conference, and player_tier, indicating varying impacts based on these categories. Figure 6, 7 and 8 show the difference in average distance from home by commitment status for the different positions, conferences and player tiers. The figures show a consistent difference between the averages for player tiers and positions. The most surprising though is top tier players. They have the largest difference among the tiers and the lowest average distance from home for committed_to being true. This is surprising because top tier players have the most choices and the opportunity to go anywhere but choose to stay closer to home. Conference differences seem inconsistent showing that the significance might be affected by interactions with the other factors.

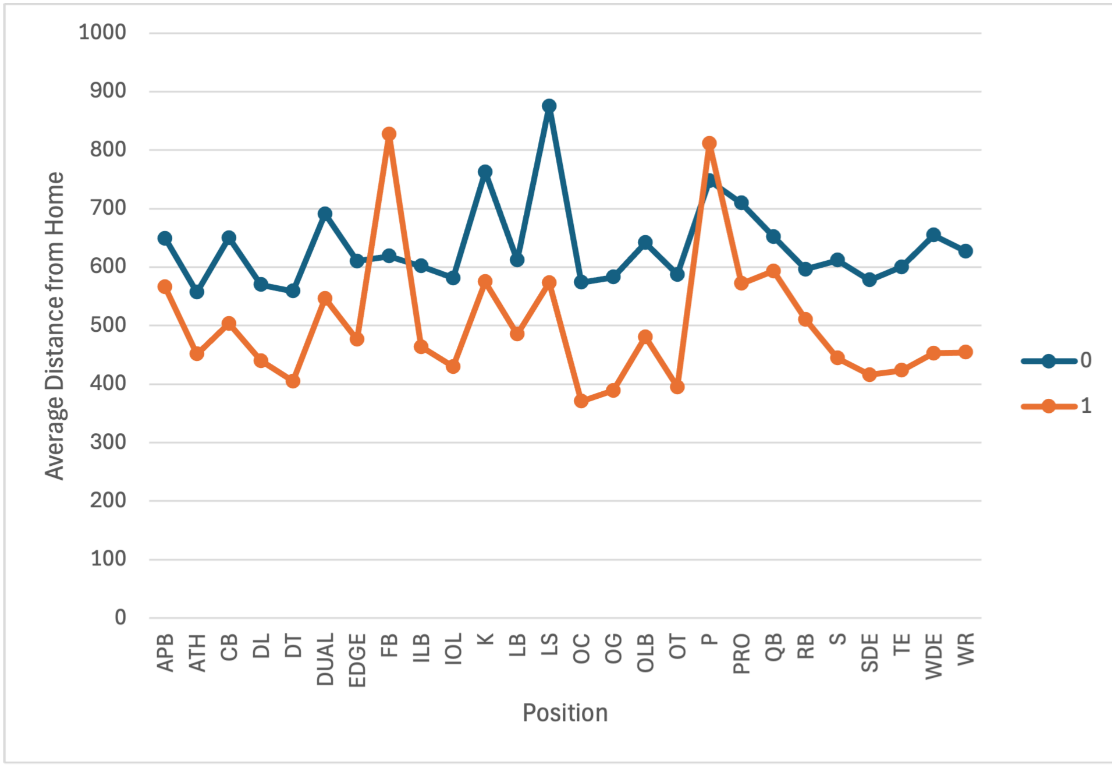

Figure 6: Average Distance From Home by Position and Commitment Status

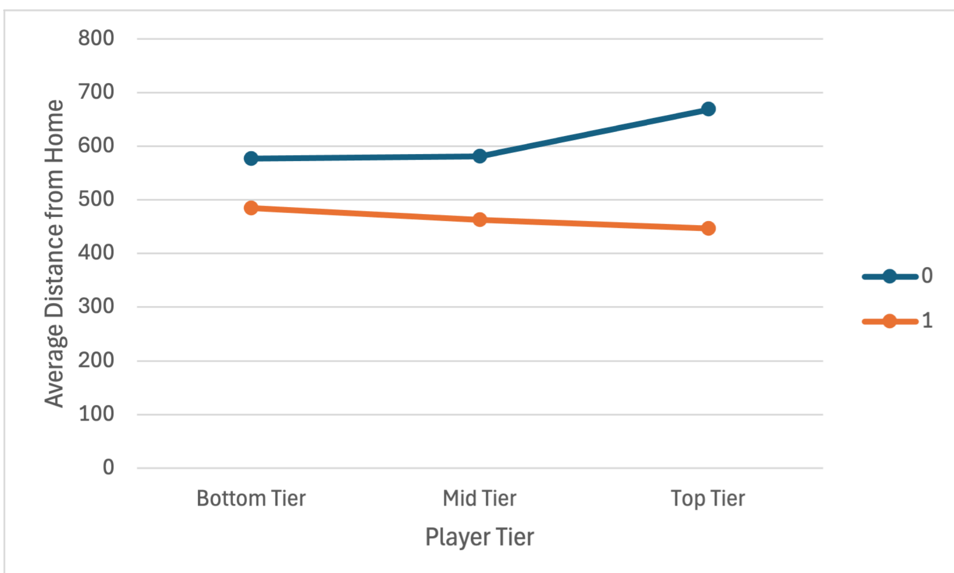

Figure 7: Average Distance From Home by Player Tier and Commitment Status

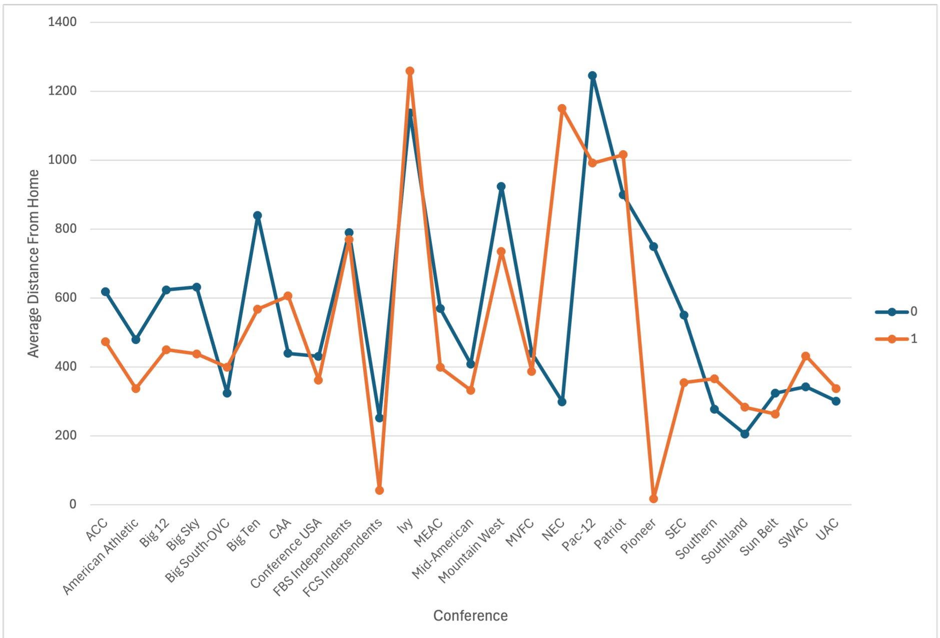

Figure 8: Average Distance From Home by Conference and Commitment Status

In view of the significant effects that were found for position, conference, and player_tier, an interaction analysis between these factors was run. To understand the significance of the effects better, three more models with different levels of interactions were run: a simplified modeled to set a base line model before more interactions were added, a model with selected interactions between conference and player_tier and a full interaction model showing all interactions between conference, player_tier, position and committed_to.

Simplified Model:

$$dist\_from\_home = C(committed\_to) + C(conference) + C(player\_tier) + C(Position) \quad (1)$$

Table 14: Simplified Model Results

	sum_sq	df	F	PR(>F)
C(committed_to)	2.470e+08	1	1005.399	8.066e-220
C(conference)	5.011e+09	24	850.0000	0.000e+00
C(player_tier)	5.464e+07	2	111.223	5.456167e-49
C(Position)	1.189e+08	25	19.354	2.540e-86
Residual	3.254e+10	132468	NaN	NaN

Model with Selected Interactions:

$$
\begin{aligned}
dist\_from\_home = C(committed\_to) + C(conference) + C(player\_tier) \\
+ C(committed\_to) * C(conference) + C(conference) * C(player\_tier)
\end{aligned}
$$

Table 15: Model with Selected Interactions Results

	sum_sq	**df**	**F**	**PR(>F)**
C(committed_to)	2.513e+08	1	1025.625	3.500e-224
C(conference)	5.010e+09	24	852.120	0.000e+00
C(player_tier)	5.976e+07	2	121.963	1.204e-53
C(committed_to):C(conference)	6.030e+07	24	10.253	1.067e-38
C(conference):C(player_tier)	1.507e+08	48	12.816	3.402e-99
Residual	3.244e+10	132421	NaN	NaN

Full Interaction:

$$
\begin{aligned}
dist\_from\_home = C(committed\_to) + C(conference) + C(player\_tier) \\
+ C(committed\_to) * C(conference) + C(conference) * C(player\_tier) \\
+ C(committed_to) * C(conference) * C(player\_tier) * C(Position)
\end{aligned}
$$

Table 16: Full Interaction Results of All Factors

	sum_sq	df	F	PR(>F)
C(committed_to)	NaN	1	NaN	NaN
C(conference)	-1.881e-01	24	-3.252e-08	1.000e+00
C(player_tier)	NaN	2	NaN	NaN
C(Position)	1.974e-02	25	3.278e-09	1.000e+00
C(committed_to): C(conference)	-5.784e-05	24	-1.000e-11	1.000e+00
C(committed_to): C(player_tier)	NaN	2	NaN	NaN
C(conference): C(player_tier)	-2.991e+08	48	-2.586e+01	1.000e+00
C(committed_to): C(Position)	-3.419e-03	25	-5.678e-10	1.000e+00
C(conference): C(Position)	3.302e+08	600	2.284e+00	1.638e-10
C(player_tier): C(Position)	1.008e+08	50	8.372e+00	2.701e-33
C(committed_to): C(conference): C(player_tier)	1.393e+07	48	1.204e+00	2.725e-01
C(committed_to): C(conference): C(Position)	1.275e+08	600	8.818e-01	8.578e-01
C(committed_to): C(player_tier): C(Position)	2.896e+07	50	2.404e+00	9.032e-02
C(conference): C(player_tier): C(Position)	5.316e+08	1200	1.839e+00	2.349e-24
C(committed_to): C(conference): C(player_tier): C(Position)	3.593e+08	1200	1.243e+00	3.545e-04
Residual	3.143e+10	130472	NaN	NaN

This model brings in all possible interactions between conference, commitment, player tier, and position. The interactions that are actually significant are conference and position, player tier and position, conference, player tier and position, and the full interaction of all factors. The full model revealed complexity and some interactions, especially involving all factors, were significant, suggesting nuanced effects across different groups.

Therefore, the significant interaction between conference and player tier suggests recruitment strategies could be optimized by considering these factors jointly. However some analyses suffered from non-normal distributions and unequal variances, impacting the robustness of the parametric tests. Co-linearity issues in interaction models necessitated model simplifications.

Distance from home significantly affects player commitment, with variability across different positions, conferences, and tiers. Recruitment strategies should consider these findings to tailor approaches to the specific needs and backgrounds of athletes.

5.1.2 Question 2: Does the quality of the stadium of a college influence a players choice? Does the position of the player effect this? Does it depend on the tier of the player?

To answer Research Question 2, the following tests were run: Two-Sample T-Tests for mean comparisons, and Chi-Square Test of Independence.

To compare the mean stadium capacities between players who committed and those who did not, without considering other factors, a Two-Sample T-Test was run. Table 17 shows the results of the test. Because the p-value < 0.05, it suggests that stadium capacity is an important factor in a player's commitment decision.

Table 17: Two-Sample T-Test

T-Stat	-18.757
p-value	0.000

To test the independence between the type of playing surface (grass) and player commitments, Chi-Square Test was run. This test was used because both committed_to and and location.grass are both categorical variables. Table 18 show the results of the test. Because the p-value < 0.05, it suggests that playing surface type is an important factor in a player's commitment decision.

Table 18: Chi-Square Test

statistic	595154915.500
p-value	4.939e-19
dof	1
expected_freq	([[83395.704, 7650.296], [37038.296, 3397.704]])

Since capacity and the type of field impacts commitment decisions, the data was divided into two more groupings to better understand further interactions. These groupings were by position and player_tier. To understand the significance of these groupings two more Two Sample T Tests were run and two more Chi-Square Tests of Independence were run. To run each Two Sample T test, the location.capacity data was group by committed_to=0 and committed_to=1 and then this was further grouped by position and player_tier. To run each Chi-Square Test of Independence, the same grouping were done but for location.grass.

Two Sample T-Tests:

By Position: Significant results were found in the majority of positions, suggesting that differences in stadium capacities influence player commitments across positions. Table 19 shows the p-value of the Two Sample T-Test for each position. Insufficient data means that the sample for commitment being true or false or both was less than 50.

Table 19: Two Sample T-Test by Position

Position	**P-value**
APB	0.016
ATH	0.000
CB	0.000
DL	0.000
DT	0.000
DUAL	0.021
EDGE	0.027
FB	Insufficient data
ILB	0.000
IOL	0.000
K	Insufficient data
LB	0.000
LS	Insufficient data
OC	0.000
OG	0.006
OLB	0.000
OT	0.000
P	Insufficient data
PRO	0.005
QB	0.054
RB	0.005
S	0.000
SDE	0.000
TE	0.000
WDE	0.087
WR	0.000

Figure 9 shows graphically the differences in average stadium capacity by position and commitment status. The largest differences are shown to be for punters, kickers and offensive center.

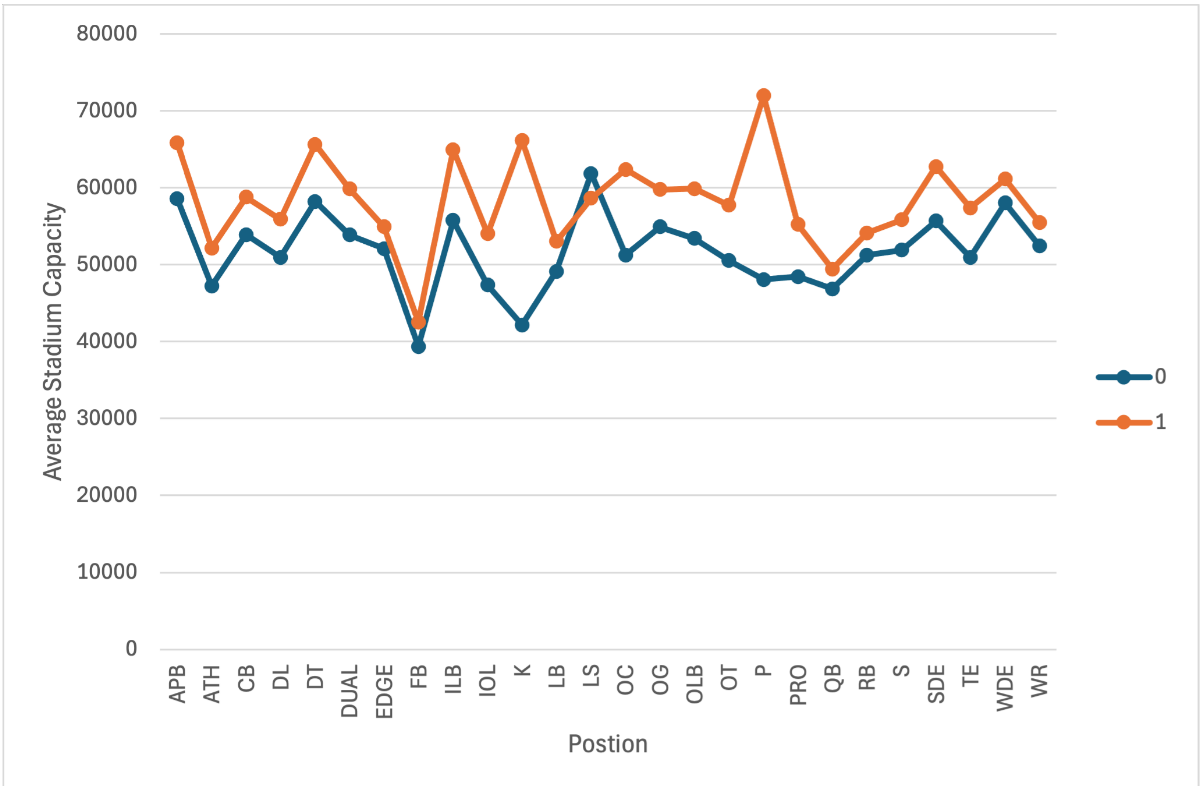

Figure 9: Average Stadium Capacity by Position and Commitment Status

By Player Tier: All player tiers showed significant differences, indicating that stadium capacity influences player commitments across different levels of player skill. Table 20 shows the p-value of the Two Sample T-Test for each position.

Table 20: Two Sample T-Test by Player Tier

Player Tier	P-value
Bottom Tier	1.877e-10
Mid Tier	2.276e-133
Top Tier	4.761e-248

Chi-Square Test of Independence:

By Position for Grass: Significant results for certain positions indicate that the type of playing surface can influence commitment decisions, particularly for positions like Kickers and Offensive Centers. Table 21 shows the p-value of the Chi-Square test for each position.

Table 21: Chi-Square Test by Position

Position	P-value
APB	0.342
ATH	0.005
CB	0.135
DL	0.072
DT	0.309
DUAL	0.071
EDGE	0.084
FB	0.876
ILB	0.010
IOL	0.051
K	4.077e-06
LB	0.248
LS	1.000
OC	0.002
OG	0.009
OLB	2.854e-05
OT	0.001
P	0.691
PRO	0.005
QB	1.000
RB	0.296
S	0.011
SDE	0.028
TE	0.129
WDE	0.200
WR	0.031

By Player Tier for Grass: Mid and Top Tiers showed significant associations, suggesting preferences for grass surfaces in higher skill tiers. Table 22 shows the p-value of the Chi-Square test for each position.

Table 22: Chi-Square Test by Player Tier

Player Tier	P-value
Bottom Tier	0.473
Mid Tier	4.047e-41
Top Tier	1.147e-32

Figures 10, 11, and 12 show as the caliber of player increases so does the percent of schools recruits commit to having grass fields.

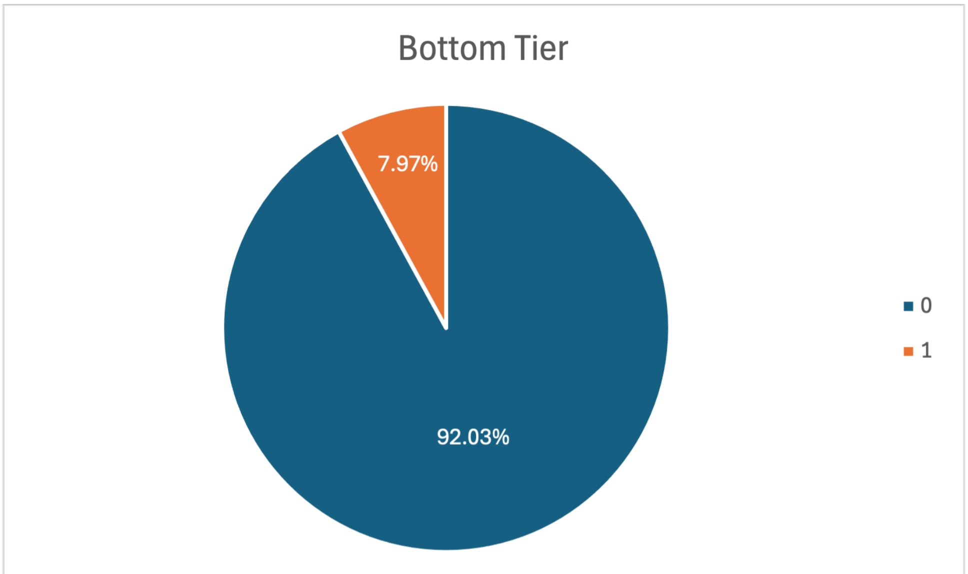

Figure 10: Percent of Commitments to Schools with Grass or Turf Fields for Bottom Tier Players

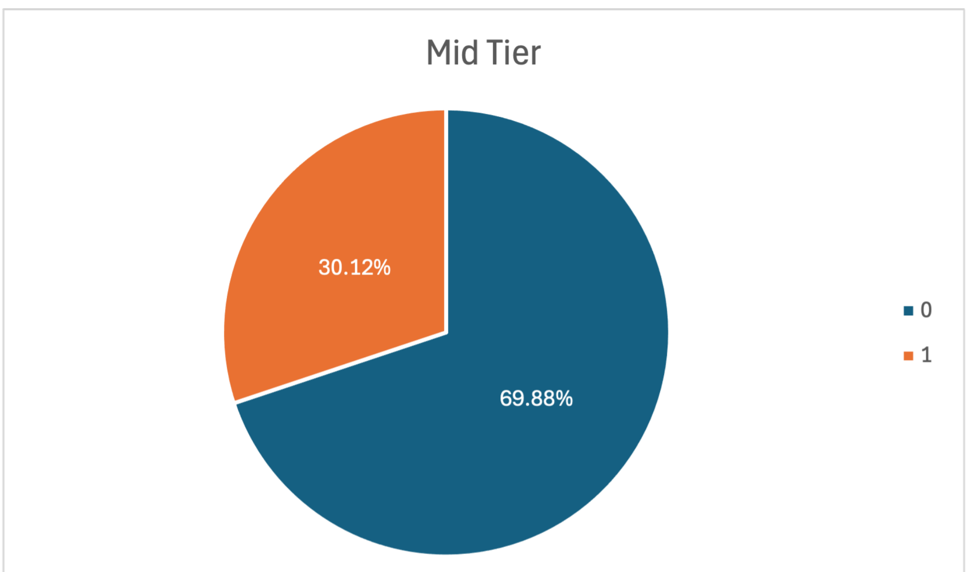

Figure 11: Percent of Commitments to Schools with Grass or Turf Fields for Mid Tier Players

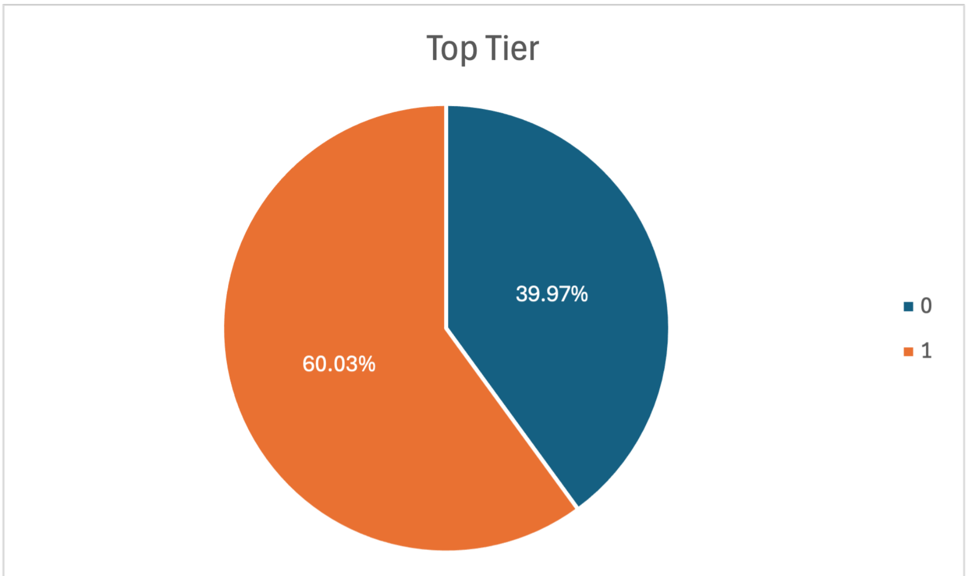

Figure 12: Percent of Commitments to Schools with Grass or Turf Fields for Top Tier Players

These results suggest that both stadium capacity and playing surface type are important factors in a player's commitment decision, influenced by both the position they play and their skill tier. This insight can guide recruitment strategies and facility investments.

5.2 Models

Before running any of the models, a correlation matrix was run to test and understand the features in the dataset. The correlation matrix provides insight into the linear relationship between variables. Values range from -1 (perfect negative correlation) to +1 (perfect positive correlation), with 0 indicating no correlation. Here are some significant findings:

A strong negative correlation (-0.888) suggests that higher composite scores typically correspond to better (lower numerical value) overall rankings. This implies that the composite score is a significant predictor of ranking, potentially useful for predictive modeling.

Similarly, Position Rank shows a high positive correlation (0.878) with Overall Ranking. Athletes with better position ranks tend to have worse overall rankings, which might seem counterintuitive and warrants further investigation into how these ranks are defined or calculated.

Both these features exhibit moderate negative correlations with Overall Ranking, indicating that athletes from schools with better records and tougher schedules tend to rank higher overall. This could reflect the competitive environment influencing an athlete's visibility and ranking.

Let's now dive into the different models run.

5.2.1 Overall

Five types of models were run and evaluated on accuracy of predicting where a recruit will commit. The model with the highest accuracy was selected. If the highest accuracy was equal across multiple models or very similar, the model with the highest accuracy and lowest log loss was selected. Table 23 shows the accuracy and log loss for each model. The five model types are logistic regression, random forest, gradient boosting, XGBoost and support vector machine.

Table 23: Comparison of Different Model Types By Accuracy, Parameters, and Log Loss

Model	Best Accuracy	Best Parameters	Log Loss
Logistic Regression	0.874	{'solver': 'lbfgs', 'C': 0.1}	0.132
Random Forest	0.915	{'n_estimators': 100, 'min_samples_split': 2, 'max_depth': None}	0.056
XGBoost	0.917	{'n_estimators': 100, 'max_depth': 6, 'learning_rate': 0.1}	0.152
SVM	0.913	{'kernel': 'rbf', 'C': 1}	0.2159
Gradient Boosting	0.913	{'n_estimators': 100, 'max_depth': 3, 'learning_rate': 0.1}	0.223

Based on the metrics of log loss and accuracy, the Random Forest model is the best model for predicting a recruit's commitment. The Random Forest model is a robust ensemble learning method that operates by building a multitude of decision trees during training and outputs the mode of the classes (classification) or mean prediction (regression) of the individual trees. Random Forests perform well for a wide range of data types and are less likely to overfit compared to some other models. Also it makes sense for the overall model because recruits are trying to make multiple choices at once.

5.2.2 By Position

Five types of models were run for each position and evaluated on accuracy of predicting where a recruit will commit based on the position they play. The model with the highest accuracy was selected. If the highest accuracy was equal across multiple models or very similar, the model with the highest accuracy and best F1 score was selected. Table 24 shows the accuracy and F1 score for each model type by position and what type of model was selected. LR is logistic regression, RF is random forest, Tuned RF is tuned random forest, GB is gradient boosting and Two-Stage is two-stage decision tree and random forest model. Figure 13 shows the frequency of each of the metrics which were selected across the best models for each position. The most frequent features were position ranking, overall ranking, composite score, in_state_1 (a recruit is in state), rank_state, dist_from_home, location.capacity, off_match_1 (went on a official visit there), unoff_match_1 (went on an unofficial visit there), and National_1 (is a National University in US News rankings).

Table 24: Performance Metrics Across Various Positions and Models

Position	Best Model Accuracy	Best Model F1 Score	Best Model
RB	0.915	0.478	LR
DT	0.907	0.539	GB
SDE	0.905	0.516	GB
EDGE	0.927	0.481	LR
WR	0.920	0.479	LR
OT	0.913	0.570	RF
IOL	0.917	0.524	GB
ATH	0.905	0.505	GB
TE	0.913	0.586	RF
LB	0.914	0.482	GB
CB	0.921	0.492	GB
DL	0.922	0.517	GB
OC	0.923	0.587	LR
QB	0.895	0.472	GB
S	0.924	0.480	LR
OG	0.923	0.498	LR
K	0.750	0.649	LR
WDE	0.924	0.480	LR
ILB	0.914	0.555	LR
OLB	0.924	0.532	LR
LS	1.000	1.000	Two-Stage RF
APB	0.900	0.474	LR
P	0.500	0.438	GB
PRO	0.879	0.468	LR
DUAL	0.900	0.502	LR
FB	1.000	1.000	Two-Stage RF

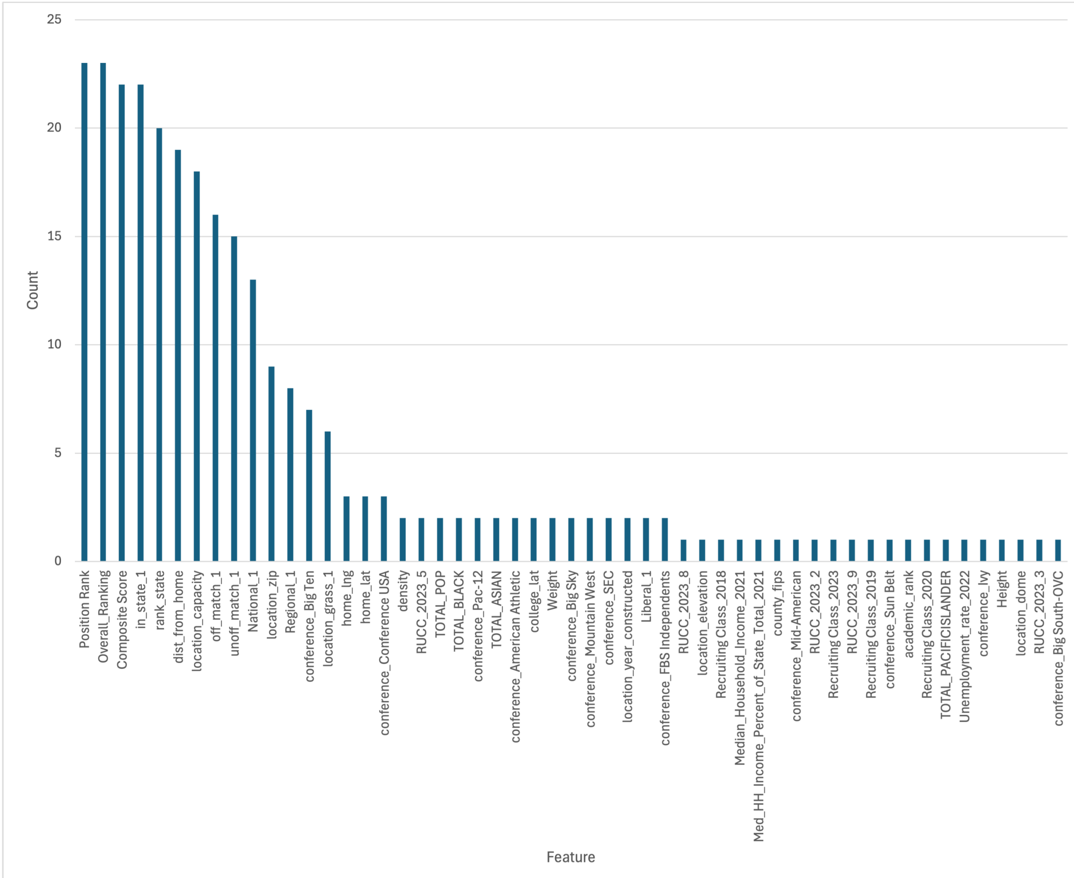

Figure 13: Count of Features Across the Position Models

Common features across models include rankings and scores, which intuitively play significant roles in a player's commitment decision. Features like "in_state_1" (whether the player is from the same state as the college) often appear, suggesting geographical proximity is a factor in commitment decisions. Logistic Regression frequently emerges as the best model, indicating that the relationship between the selected features and the target might be linear. However, where complex patterns exist (e.g., for long snappers), tuned models like the Tuned Random Forest perform better, highlighting the need for more sophisticated approaches in those cases. The models generally show high accuracy, suggesting they are good at predicting outcomes. However, the precision and recall for the minority class (likely the players who commit) are often low, indicating that the models may struggle to correctly identify the less frequent outcome of player commitment. This could be due to class imbalance where the number of non-commitments far exceeds commitments. The poor performance on the minority class suggests that there might be an imbalance in the dataset between players who commit and those who do not. This can lead to models that are biased towards predicting non-commitments. Certain features consistently appear across models for different positions, suggesting they have strong predictive power. These include player rankings, scores, and state-related features.

5.2.3 By Player Tier

Five types of models were run for each player tier and evaluated on accuracy of predicting where a recruit will commit based on the tier of player they are. The model with the highest accuracy was selected. If the highest accuracy was equal across multiple models or very similar, the model with the highest accuracy and best F score was selected. Table 25 shows the accuracy and F score for each model type by position and what type of model was selected. LR is logistic regression, RF is random forest, Tuned RF is tuned random forest, GB is gradient boosting and Two-Stage is two-stage decision tree and random forest model. Figure 14 shows the frequency of each of the metrics which were selected across the best models for each position. The metrics seen in multiple player tier models were: location.capacity, dist_from_home, in_state_1, location.zip, Regional_1 (is a Regional University in US News rankings) and National_1.

Table 25: Performance Metrics Across Various Player Tiers and Models

Player Tier	Best Model Accuracy	Best Model F1 Score	Best Model
Bottom Tier	0.807	0.531	GB
Mid Tier	0.909	0.565	RF
Top Tier	0.948	0.512	GB

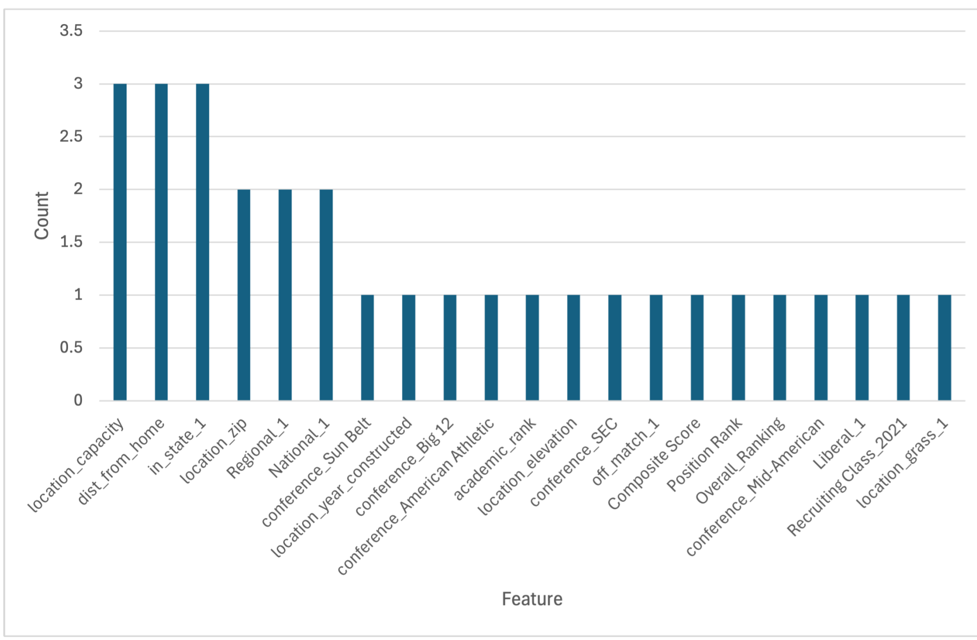

Figure 14: Count of Features Across the Player Tier Models

For the Bottom Tier, Gradient Boosting performed best with an accuracy of 80.7%, primarily using features like in_state_1 and Regional_1. Key Features: Included binary variables like in_state_1 and Regional_1, suggesting that geographical factors play a significant role in the classification at this tier. Shows a high precision for the majority class (0.0) but a complete miss for the minority class (1.0), indicating that the model could not identify any of the positive class correctly. This is a classic case of class imbalance impacting model performance.

In the Mid Tier, Random Forest had the highest accuracy at 90.9%, with significant contributions from features like location_capacity and dist_from_home. location_capacity and dist_from_home were important, highlighting that logistics and proximity to home influence player commitments at this level. Much like the Bottom Tier, the model excels at predicting the majority class but struggles significantly with the minority class, which can be detrimental if those predictions are crucial to decision-making.

The Top Tier also saw Gradient Boosting as the best model, with an accuracy of 94.8%, where location_capacity was the most impactful feature. Dominated by location_capacity, suggesting that for top-tier players, the size and capacity of the venue play a significant role in their decisions. Exhibits a similar pattern to the other tiers, with excellent performance on the majority class but very poor recall on the minority class.

5.3 Conclusion

Through both the models and the research questions, it is clear that geographical location, rank or tier of a player, and position matter most in recruitment. Therefore if a team is looking at a high school recruit, minus any financial data, that team should focus on those recruits closer to the school and his caliber.

6 Future Works

Now that some of the metrics that predict a recruit's choice of school are more clear, other aspects of the sport can be brought into the analysis. Future work includes incorporating financial data and the transfer portal.

On July 1, 2021, the NCAA implemented a policy where athletes can benefit off the use of their name, image and likeness (NIL) [33]. Due to this, recruits can now negotiate the equivalent of a contract and earn more money from certain schools over others. This financial data would be a great addition to the model for future analysis. It was not included in this analysis because this information is generally not publicly available and does not have to be disclosed. Once, this information is accessible, it could be used to analyze the impact of a player's tier, position, and conference on NIL money and if it is significant in whether or not a player commits to a school.

Around the same time as the start of NIL, in April 2021, the NCAA made changes to the transfer portal rules [13]. At this time, the NCAA allowed for a one-time transfer for all college athletes without the need to sit out. Before, an athlete had to sit out a season. Then, in April 2024, the NCAA changed the rules again allowing for unlimited transfers without the need to sit out [14]. Now, colleges aren't just recruiting high school athletes. Colleges compete on athletes entering the portal and whether they can achieve their needs from the portal versus a high school athlete. Future work would include a model for predicting where high school athletes commit, a model for predicting where college athletes will transfer, and a combined model that produces the best athletes both from the high school pool and the college pool for a specific school's needs. This final model could evaluate based off of NIL expectations, position and expected play.

References

[1] Academicinfo.net. (n.d.). College location: Close to home or far away? Retrieved from https://www.academicinfo.net/college-prep/college-location-close-to-home-or-far-away

[2] Alceo, P., & Henriques, R. (2019). Sports analytics: Maximizing precision in predicting MLB base hits. Proceedings of the 11th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management, 190–201. doi:10.5220/0008362201900201

[3] Almujahed, S., Ongor, N., Tigmom, J., Sagoo, N., Mantialla, J., Chao, F., & Kendrick, P. (2013). Sports analytics: Designing a decision-support tool for game analysis using big data. George Mason University.

[4] Anani, Z. (2025, January 28). The true size of the global sports industry. GIS. Retrieved from https://gis.sport/news/the-true-size-of-the-global-sports-industry/

[5] Badenhausen, K. (2024). How NFL teams and owners make their money. Retrieved from https://www.sportico.com/leagues/football/2024/how-nfl-teams-owners-make-money-1234795113/

[6] Baumer, B. S., Matthews, G. J., & Nguyen, Q. (2023). Big ideas in sports analytics and statistical tools for their investigation. WIREs Computational Statistics, 15(6), e1612. https://doi.org/10.1002/wics.1612

[7] Blanchfield, J. E., Hargroves, M. T., Keith, P. J., Lansing, M. C., Nordin, L. H., Palmer, R. C., … Napoli, N. J. (2019). Developing Predictive Athletic Performance Models for Informative Training Regimens. 2019 Systems and Information Engineering Design Symposium (SIEDS), 1–6. doi:10.1109/SIEDS.2019.8735633

[8] Brock Purdy stats, height, weight, position, draft, college. (2025a). Retrieved from https://www.pro-football-reference.com/players/P/PurdBr00.htm

[9] Brooks, A. (2024). ESPN delivers record viewership across College Football Playoff and New Year's six. Retrieved from https://espnpressroom.com/us/press-releases/2024/01/espn-delivers-record-viewership-across-college-football-playoff-and-new-years-six/#: :text=National%20Championship%20Notches%20Multi%2DYear, Spartanburg%2DAsheville%20(19.2).

[10] Coleman, B. (2012). Identifying the "Players" in sports analytics research. Interfaces, 42(2), 109-118.

[11] Davis, J., Bransen, L., Devos, L., et al. (2024). Methodology and evaluation in sports analytics: Challenges, approaches, and lessons learned. Machine Learning, 113, 6977–7010. https://doi.org/10.1007/s10994-024-06585-0

[12] De la Torre, R., Calvet, L. O., Lopez-Lopez, D., Juan, A. A., & Hatami, S. (2022). Business analytics in sport talent acquisition: Methods, experiences, and open research opportunities. International Journal of Business Analytics, 9(1), 1-20. https://doi.org/10.4018/IJBAN.290406

[13] Dellenger, R. (2021, April 14). "It's going to change the landscape": The NCAA's

transfer revolution is here. SI. Retrieved from https://www.si.com/college/2021/04/14/ncaa-transfers-rule-change-football-basketball

[14] Durham Wright, M. D. (2024, April 17). Division I Council approves changes to transfer rules. NCAA.org. Retrieved from https://www.ncaa.org/news/2024/4/17/media-center-division-i-council-approves-changes-to-transfer-rules.aspx

[15] Elkins, H., et al. (2017). Implementing data analytics for U.Va. football. Proceedings of the 2017 Systems and Information Engineering Design Symposium (SIEDS), 202-207. https://doi.org/10.1109/SIEDS.2017.7937717

[16] Fortune Business Insights. (n.d.). Sports analytics market size, share: Growth analysis [2032]. Retrieved from https://www.fortunebusinessinsights.com/sports-analytics-market-102217

[17] Friesen, K. B., Barfield, J. W., Murrah, W. M., Dugas, J. R., Andrews, J. R., & Oliver, G. D. (2021). The association of upper-body kinematics and earned run average of NCAA Division I softball pitchers. Journal of Strength and Conditioning Research, 35(11), 3145-3150. https://doi.org/10.1519/JSC.0000000000003287

[18] Frost, W., Groom, R., & Nicholls, S. B. (2025). The use of performance analysis and data-driven approaches within senior men's football recruitment. International Journal of Sports Science & Coaching, 20(2), 604-616. https://doi.org/10.1177/17479541251315948

[19] Fury, M., Scarborough, D., Oh, L., Wright-Chisem, J., Fury, J., & Berkson, E. (2021). Performance analytics and pitch metrics as predictors of ulnar collateral ligament injury in Major League Baseball pitchers. Orthopaedic Journal of Sports Medicine, 9(10_suppl5). https://doi.org/10.1177/2325967121S00271

[20] Ghosh, I., Ramamurthy, S. R., Chakma, A., & Roy, N. (2023). Sports analytics review: AI applications, emerging technologies, and algorithmic perspectives. WIREs Data Mining and Knowledge Discovery, 13(5), e1496. https://doi.org/10.1002/widm.1496

[21] Gorman, D. (2017). Sports analytics: Analysis of the NFL (BSc thesis). National College of Ireland.

[22] Gowda, M., et al. (2017). Bringing IoT to sports analytics. USENIX Symposium, 499–513.

[23] Herberger, T. A., & Litke, C. (2021). The impact of big data and sports analytics on professional football. In Digitalization, Digital Transformation and Sustainability in the Global Economy (pp. 147-162). Springer. https://doi.org/10.1007/978-3-030-77340-3_12

[24] Hu, Y., & Alturjman, S. (2022). Timing system of track and field competition based on data analysis algorithms. Springer Lecture Notes, vol 123, 48-56. https://doi.org/10.1007/978-3-030-96908-0_6

[25] Kaskenmaa, M. (2023). Using data analytics in hockey player talent identification (Master's thesis). Oulu University.

[26] Kiddy, C. (2025). Ranking the biggest & smallest college football (FBS) stadiums. Retrieved from https://sports.betmgm.com/en/blog/college-football/ranking-

biggest-smallest-college-football-stadiums-fbs-ncaaf-bm06-2/

[27] Kniffin, K. M., Howley, T., & Bardreau, C. (2017). Putting muscle into sports analytics: Strength, conditioning, and ice hockey performance. Journal of Strength and Conditioning Research, 31(12), 3253-3259. https://doi.org/10.1519/JSC.0000000000002211

[28] Loughran GJ, Vulpis CT, Murphy JP, et al. Incidence of Knee Injuries on Artificial Turf Versus Natural Grass in National Collegiate Athletic Association American Football: 2004-2005 Through 2013-2014 Seasons. The American Journal of Sports Medicine. 2019;47(6):1294-1301. doi:10.1177/0363546519833925

[29] Meyers, B. R., Burns, M., Coughlin, B. Q., & Bolte, E. (2021). Expected goals model for lacrosse. The Sport Journal. Retrieved from https://thesportjournal.org/article/on-the-development-and-application-of-an-expected-goals-model-for-lacrosse/

[30] Mizels, J., Erickson, B., & Chalmers, P. (2022). Data and analytics in baseball. Current Reviews in Musculoskeletal Medicine, 15, 283-290. https://doi.org/10.1007/s12178-022-09763-6

[31] Moorefield, J. B. (2021). Oakland Athletics' use of sabermetrics and big data analytics (Honors Thesis). University of Tennessee.

[32] Morgulev, E., Azar, O. H., & Lidor, R. (2018). Sports analytics and the big-data era. International Journal of Data Science and Analytics, 5, 213–222. https://doi.org/10.1007/s41060-017-0093-7

[33] Palomba, A. (2024, November 20). Understanding the NCAA transfer portal and recent rule changes. SportsRecruits Blog. Retrieved from https://blog.sportsrecruits.com/2024/04/30/understanding-the-ncaa-transfer-portal-and-recent-rule-changes/

[34] Pantzalis, V. C., & Tjortjis, C. (2020). Sports analytics for football league table and player performance prediction. 2020 11th International Conference on Information, Intelligence, Systems and Applications (IISA), 1-8. https://doi.org/10.1109/IISA50023.2020.9284352

[35] Passfield, L., & Hopker, J. G. (2017). A mine of information: Can sports analytics provide wisdom from your data? International Journal of Sports Physiology and Performance, 12(7), 851-855. https://doi.org/10.1123/ijspp.2016-0644

[36] Pelechrinis, K. (2023). Winning in American Football with Data and Analytics. In Statistics Meets Sports: What We Can Learn from Sports Data (pp. 43–74). essay, Newcastle upon Tyne: Cambridge Scholars Publishing.

[37] Pinger, N. (2015). Home field advantage: The facts and the fiction. Retrieved from https://www.chicagobooth.edu/review/home-field-advantage-facts-and-fiction

[38] Press, A. (2024). NCAA generates nearly $1.3 billion in revenue for 2022-23. Retrieved from https://www.espn.com/college-sports/story/_/id/39439274/ncaa-generates-nearly-13-billion-revenue-2022-23

[39] Press, A. (2025). Super Bowl LIX averages record audience of 127.7 million viewers. Retrieved from

https://www.nfl.com/news/super-bowl-lix-averages-record-audience-of-127-7-million-viewers#: :text=Despite%20the%20game%20being%20a,8:15%20p.m.%20ET).

[40] Radke, D., Radke, D., Brecht, T., & Pawelczyk, A. (2021). Passing and pressure metrics in ice hockey. In Artificial Intelligence for Sports Analytics (AISA) Workshop at IJCAI '21. Retrieved from https://www.researchgate.net/publication/352465415_Passing_and_Pressure_Metrics_in_Ice_Hockey

[41] Rees, L., Rakes, T., & Deane, J. (11 2015). Using Analytics To Challenge Conventional Baseball Wisdom. Journal of Service Science (JSS), 8, 11. doi:10.19030/jss.v8i1.949

[42] Reid, W. (2018a). 'Hooball: Students help UVA Football Gain Data Analytics Edge. Retrieved from https://news.virginia.edu/content/hooball-students-help-uva-football-gain-data-analytics-edge

[43] Roumani, Y. F. (2023). Sports analytics in the NFL: Classifying the winner of the Superbowl. Annals of Operations Research, 325, 715–730. https://doi.org/10.1007/s10479-022-05063-x

[44] Sabin, R. (2021). Estimating player value in American football using plus-minus models. Journal of Quantitative Analysis in Sports, 17(4), 313-364. https://doi.org/10.1515/jqas-2020-0033

[45] Sarlis, V., & Tjortjis, C. (2020). Sports analytics—evaluation of basketball players and team performance. Information Systems, 93, 101562.

[46] Sarlis, V., Chatziilias, V., Tjortjis, C., & Mandalidis, D. (2021). A data science approach analyzing the impact of injuries on basketball player and team performance. Information Systems, 99, 101750.

[47] Sarlis, V., Papageorgiou, G., & Tjortjis, C. (2024). Leveraging sports analytics and association rule mining to uncover recovery and economic impacts in NBA basketball. Data, 9(7), 83. https://doi.org/10.3390/data9070083

[48] Schoborg, C. (2023). Football by the numbers: A look into sports analytics currently used in the National Football League (Master's thesis). University of Central Florida. Retrieved from https://stars.library.ucf.edu/etd2020/1762

[49] SportsRecruits. (n.d.). Build your target list: Identifying fit, reach, and safety schools. Retrieved from https://sportsrecruits.com/resources/how-to-get-recruited/building-a-target-list

[50] Steinberg, L. (2015, August 18). Changing the game: The rise of sports analytics. Forbes. Retrieved from https://www.forbes.com/sites/leighsteinberg/2015/08/18/changing-the-game-the-rise-of-sports-analytics/

[51] Terner, Z., & Franks, A. (2021). Modeling player and team performance in basketball. Annual Review of Statistics and Its Application, 8, 1-23.

[52] Tom Brady stats, height, weight, position, draft, college. (2025). Retrieved from https://www.pro-football-reference.com/players/B/BradTo00.htm

[53] Wang, W. W., Johnston, R., & Jones, K. (09 2011). Home Advantage in American College Football Games: A Multilevel Modelling Approach. Journal of Quantitative Analysis in Sports, 7, 23–23. doi:10.2202/1559-0410.1328

Appendix A.1: Data Dictionaries

A.1.1: Overall Data Dictionary

Table 26: Description of Features

Feature	Definition
247Sports ID	Recruits ID on 247 Sports
row_count	Number of rows in dataset of recruit
First Name	First Name of Recruit
Last Name	Last Name of Recruit
Overall Ranking	The ranking of the recruit on 247Sports
Position	Position recruits plays
Height	Height of the Recruit in inches
Weight	Weight of the Recruit in pounds
High School	High School where recruit attended
City	City where college is
initials_state	Initials of State where college is
Commitment Status	College that Recruit Committed To
Composite Score	Score of recruit calculated by 247Sports algorithm; combined data from all four recruiting services into an overall score, thus reducing the potential randomness of focusing on ratings from a single recruiting service.
Position Rank	The ranking of the recruit on 247Sports by position
rank_state	The ranking of the recruit on 247Sports by state
Recruiting Class	Year in which recruit was recruited
official_visits	Number of official visits to colleges
college_name	Name of the college
college_abbreviation	Abbreviation of College name
conference	Conference of school
classification	if FBS or FCS school
location.name	Stadium of college name
location.city	Stadium of college city
location.state	Stadium of college state
location.zip	Stadium of college zip code
location.timezone	Stadium of college timezone
college_lat	College's latitude
college_long	College's longitude
location.elevation	Stadium of college elevation
location.capacity	Stadium of college capacity
location.year_constructed	Stadium of college year constructed
location.grass	If a stadium has a grass field or not
location.dome	If a stadium has a dome field or not
National	If school is a national school on US News
Liberal	If school is a liberal school on US News

Regional	If school is a regional school on US News
Record %	Win record of a college over the last 5 years
academic_rank	Rank of school on US News
dist_from_home	Distance a college is from a recruit's in miles
in_state	If a recruit is in state or not
player_tier	Tier of player calculated from composite score

A.1.2: Categorical Data Dictionaries

Table 27: Positions Data Dictionary

Position	Definition
APB	All Purpose Back
ATH	Athlete
CB	Cornerback
DL	Defensive Line
DT	Defensive Tackle
DUAL	Dual-Threat Quarterback
EDGE	Edge
FB	Fullback
ILB	Inside Linebacker
IOL	Interior Offensive Lineman
K	Kicker
LB	Linebacker
LS	Long Snapper
OC	Center
OG	Offensive Guard
OLB	Outside Linebacker
OT	Offensive Tackle
P	Punter
PRO	Pro-Style Quarterback
QB	Quarterback
RB	Running back
S	Safety
SDE	Strong-Side Defensive End
TE	Tight End
WDE	Wide-Side Defensive End
WR	Wide Receiver

Table 28: Player Tier Data Dictionary

Player Tier	Definition
Bottom Tier	Composite Score between 0.7 and 0.8, including 0.7
Mid Tier	Composite Score between 0.8 and 0.9, including 0.8
Top Tier	Composite Score between 0.9 and 1.0, including both

Table 29: Rural Urban Continuum Code Data Dictionary

Rural Urban Continuum Code	Definition
1	Total population of the metro area: 1 million people or more
2	Total population of the metro area: 250,000 to 1 million people
3	Total population of the metro area: below 250,000
4	Nonmetro counties with urban population of 20,000 or more
5	Nonmetro counties with urban population of 20,000 or more
6	Nonmetro counties with urban population of 5,000 to 20,000
7	Nonmetro counties with urban population of 5,000 to 20,000
8	Nonmetro counties with urban population of fewer than 5,000
9	Nonmetro counties with urban population of fewer than 5,000

Appendix A.2: Additional Tables

Table 30: Average Composite Score by Position

Position	Average Composite Score
APB	0.898
ATH	0.872
CB	0.888
DL	0.883
DT	0.888
DUAL	0.895
EDGE	0.893
FB	0.825
ILB	0.888
IOL	0.874
K	0.818
LB	0.879
LS	0.800
OC	0.883
OG	0.879
OLB	0.883
OT	0.881
P	0.816
PRO	0.887
QB	0.890
RB	0.885
S	0.883
SDE	0.880
TE	0.877
WDE	0.892
WR	0.890

Table 31: Average Composite Score by Conference

Conference	Average Composite Score
ACC	0.898
American Athletic	0.863
Big 12	0.891
Big Sky	0.839
Big South-OVC	0.841
Big Ten	0.904
CAA	0.841
Conference USA	0.859
FBS Independents	0.891
FCS Independents	0.831
Ivy	0.847
MEAC	0.853
Mid-American	0.857
Mountain West	0.857
MVFC	0.843
NEC	0.823
Pac-12	0.882
Patriot	0.834
Pioneer	0.816
SEC	0.913
Southern	0.836
Southland	0.840
Sun Belt	0.860
SWAC	0.865
UAC	0.848

Position	Percent Uncommitted	Percent Committed	Grand Total
APB	59.09%	40.91%	100.00%
ATH	70.18%	29.82%	100.00%
CB	64.85%	35.15%	100.00%
DL	66.39%	33.61%	100.00%
DT	60.08%	39.92%	100.00%
DUAL	56.88%	43.12%	100.00%
EDGE	65.75%	34.25%	100.00%
FB	100.00%	0.00%	100.00%
ILB	56.11%	43.89%	100.00%
IOL	67.77%	32.23%	100.00%
K	51.06%	48.94%	100.00%
LB	69.70%	30.30%	100.00%
LS	60.00%	40.00%	100.00%
OC	55.29%	44.71%	100.00%
OG	58.87%	41.13%	100.00%
OLB	58.42%	41.58%	100.00%
OT	64.02%	35.98%	100.00%
P	54.55%	45.45%	100.00%
PRO	64.14%	35.86%	100.00%
QB	73.35%	26.65%	100.00%
RB	68.85%	31.15%	100.00%
S	64.80%	35.20%	100.00%
SDE	58.96%	41.04%	100.00%
TE	66.15%	33.85%	100.00%
WDE	60.00%	40.00%	100.00%
WR	66.21%	33.79%	100.00%

Table 32: Distribution of Committed Positions

Player Tier	Uncommitted	Committed	Grand Total
Bottom Tier	92.03%	7.97%	100.00%
Mid Tier	69.88%	30.12%	100.00%
Top Tier	39.97%	60.03%	100.00%
Grand Total	65.56%	34.44%	100.00%

Table 33: Table showing percentages of committed_to across different tiers

Grass or Turf Field	Uncommitted	Committed	Grand Total
0 (Turf)	84695	36724	121419
1 (Grass)	7278	3824	11102

Table 34: Count of location.grass by Commitment Status

Position	Count
APB	836
ATH	8662
CB	13274
DL	9480
DT	2924
DUAL	1102
EDGE	6118
FB	20
ILB	2085
IOL	6266
K	184
LB	8115
LS	18
OC	944
OG	2349
OLB	3086
OT	11737
P	37
PRO	1379
QB	4330
RB	8618
S	11755
SDE	2345
TE	6858
WDE	2767
WR	17232

Table 35: Counts of Each Position in Dataset

Position	**Average Weight**
APB	183.65
ATH	189.56
CB	175.85
DL	270.28
DT	290.38
DUAL	199.05
EDGE	229.28
FB	232.25
ILB	222.71
IOL	294.87
K	180.23
LB	213.05
LS	223.33
OC	289.92
OG	298.75
OLB	214.45
OT	291.90
P	189.19
PRO	201.18
QB	198.62
RB	195.94
S	186.34
SDE	254.32
TE	228.32
WDE	231.11
WR	183.98

Table 36: Average Weight by Position

Conference	Average Median Household Income (2021)
ACC	71418.572
American Athletic	69469.449
Big 12	71691.800
Big Sky	81916.568
Big South-OVC	67279.077
Big Ten	72641.803
CAA	77359.3101
Conference USA	67410.506
FBS Independent	74015.211
FCS Independent	85205.818
Ivy	76108.129
MEAC	72521.218
Mid-American	69709.788
Mountain West	78054.196
MVFC	72036.460
NEC	73184.237
Pac-12	78214.881
Patriot	78645.950
Pioneer	79763.370
SEC	68589.645
Southern	66650.889
Southland	69950.479
Sun Belt	65975.683
SWAC	65453.899
UAC	67024.668

Table 37: Average Median Household Income by Conference (2021)

Conference	**Average Record %**
ACC	0.521
American Athletic	0.484
Big 12	0.505
Big Sky	0.495
Big South-OVC	0.429
Big Ten	0.561
CAA	0.477
Conference USA	0.522
FBS Independents	0.453
FCS Independents	0.379
Ivy	0.605
MEAC	0.368
Mid-American	0.470
Mountain West	0.488
MVFC	0.514
NEC	0.356
Pac-12	0.4842
Patriot	0.475
Pioneer	0.450
SEC	0.593
Southern	0.495
Southland	0.400
Sun Belt	0.515
SWAC	0.516
UAC	0.504

Table 38: Average Record Percentages by Conference

Position	**Average Height**
APB	69.66
ATH	72.56
CB	71.92
DL	75.50
DT	74.78
DUAL	73.94
EDGE	75.55
FB	73.75
ILB	73.49
IOL	76.03
K	72.65
LB	73.75
LS	72.50
OC	75.26
OG	75.75
OLB	73.88
OT	77.63
P	74.59
PRO	74.97
QB	74.13
RB	70.76
S	72.63
SDE	75.97
TE	76.52
WDE	75.64
WR	73.06

Table 39: Average Height by Position

Conference	Uncommitted	Committed	Grand Total
ACC	90.80%	9.20%	100.00%
American Athletic	91.92%	8.08%	100.00%
Big 12	91.17%	8.83%	100.00%
Big Sky	91.78%	8.22%	100.00%
Big South-OVC	97.90%	2.10%	100.00%
Big Ten	89.99%	10.01%	100.00%
CAA	94.99%	5.01%	100.00%
Conference USA	92.61%	7.39%	100.00%
FBS Independents	90.68%	9.32%	100.00%
FCS Independents	97.73%	2.27%	100.00%
Ivy	97.57%	2.43%	100.00%
MEAC	96.68%	3.32%	100.00%
Mid-American	92.46%	7.54%	100.00%
Mountain West	90.29%	9.71%	100.00%
MVFC	96.54%	3.46%	100.00%
NEC	97.50%	2.50%	100.00%
Pac-12	90.38%	9.62%	100.00%
Patriot	98.37%	1.63%	100.00%
Pioneer	99.39%	0.61%	100.00%
SEC	90.77%	9.23%	100.00%
Southern	96.14%	3.86%	100.00%
Southland	96.08%	3.92%	100.00%
Sun Belt	92.03%	7.97%	100.00%
SWAC	97.26%	2.74%	100.00%
UAC	96.68%	3.32%	100.00%

Table 40: Percent of Commitment Status by Conference

Player Tier	Count of Grass Fields
Bottom Tier	226
Mid Tier	21845
Top Tier	18477

Table 41: Count of Grass Fields by Player Tier

Player Tier	Count
Bottom Tier	2498
Mid Tier	90621
Top Tier	39402

Table 42: Distribution of Players by Tier

Player Tier	Count of Players In State
Bottom Tier	579
Mid Tier	17699
Top Tier	5845

Table 43: Count of In State Attendance by Tier

Rural Urban Continuum Code	Average of Median Household Income (2021)
1	76032.845
2	62501.445
3	58697.311
4	55011.128
5	51688.853
6	49566.830
7	47011.644
8	52046.307
9	47124.993

Table 44: Average of Median Household Income 2021 by Rural Urban Continuum Code

National University	Count
0 (False)	97632
1 (True)	21450

Table 45: Counts of National Universities

Table 46: Sum of In-State Recruits by State

State	Sum of In-State Recruits
AL	905

AR	174
AZ	242
CA	1371
CO	105
DC	7
DE	1
FL	3042
GA	1121
HI	34
IA	139
ID	26
IL	341
IN	355
KS	73
KY	184
LA	1031
MA	40
MD	165
MI	520
MN	31
MO	61
MS	384
MT	12
NC	952
ND	8
NE	42
NJ	174
NM	20
NV	80
NY	84
OH	1145
OK	195
OR	53
PA	433
SC	253
SD	9
TN	846
TX	8499
UT	181
VA	599
WA	123
WI	30

WV	32
WY	1

Conference	Percent Out of State	Percent In State	Grand Total
ACC	81.85%	18.15%	100.00%
American Athletic	67.76%	32.24%	100.00%
Big 12	76.70%	23.30%	100.00%
Big Sky	71.73%	28.27%	100.00%
Big South-OVC	72.34%	27.66%	100.00%
Big Ten	92.62%	7.38%	100.00%
CAA	81.44%	18.56%	100.00%
Conference USA	71.86%	28.14%	100.00%
FBS Independents	97.87%	2.13%	100.00%
FCS Independents	97.73%	2.27%	100.00%
Ivy	97.89%	2.11%	100.00%
MEAC	60.63%	39.37%	100.00%
Mid-American	79.67%	20.33%	100.00%
Mountain West	86.45%	13.55%	100.00%
MVFC	85.54%	14.46%	100.00%
NEC	56.25%	43.75%	100.00%
Pac-12	95.97%	4.03%	100.00%
Patriot	94.12%	5.88%	100.00%
Pioneer	78.79%	21.21%	100.00%
SEC	88.91%	11.09%	100.00%
Southern	73.67%	26.33%	100.00%
Southland	19.67%	80.33%	100.00%
Sun Belt	74.14%	25.86%	100.00%
SWAC	60.18%	39.82%	100.00%
UAC	70.48%	29.52%	100.00%

Table 47: Percent of In-State and Out-of-State by Conference

Appendix A.3: Additional Graphs

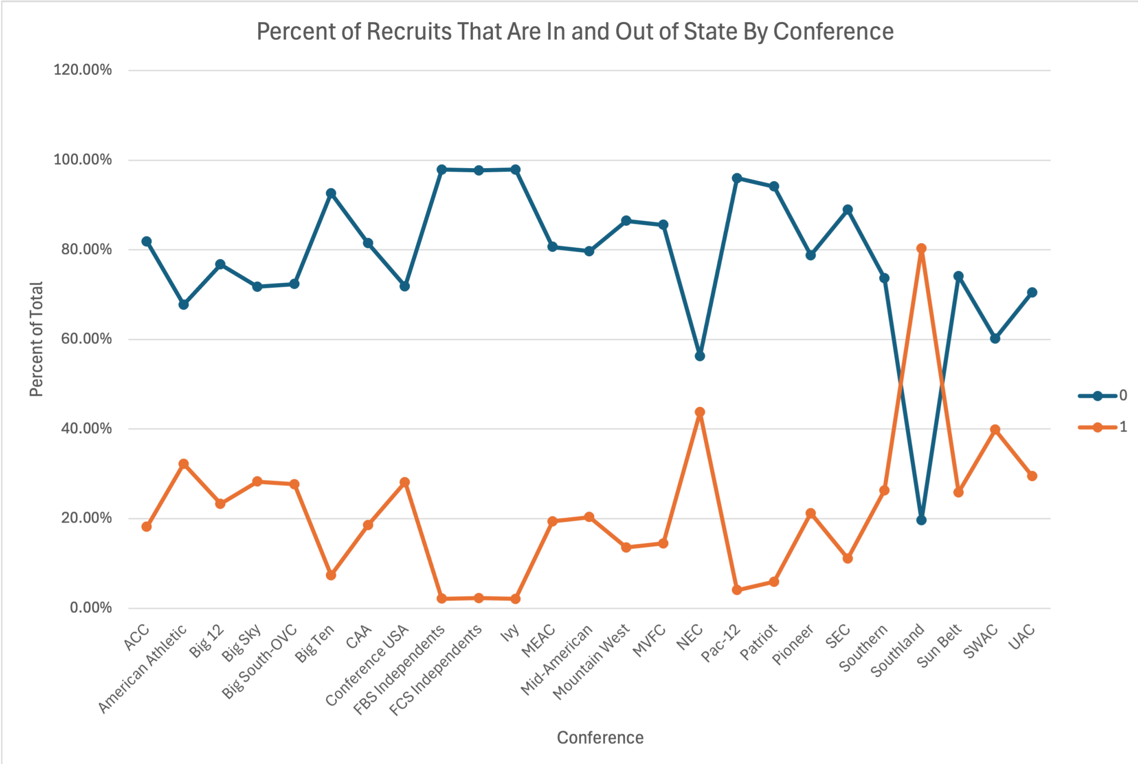

Figure 15: In and Out of State Percentage by Conference

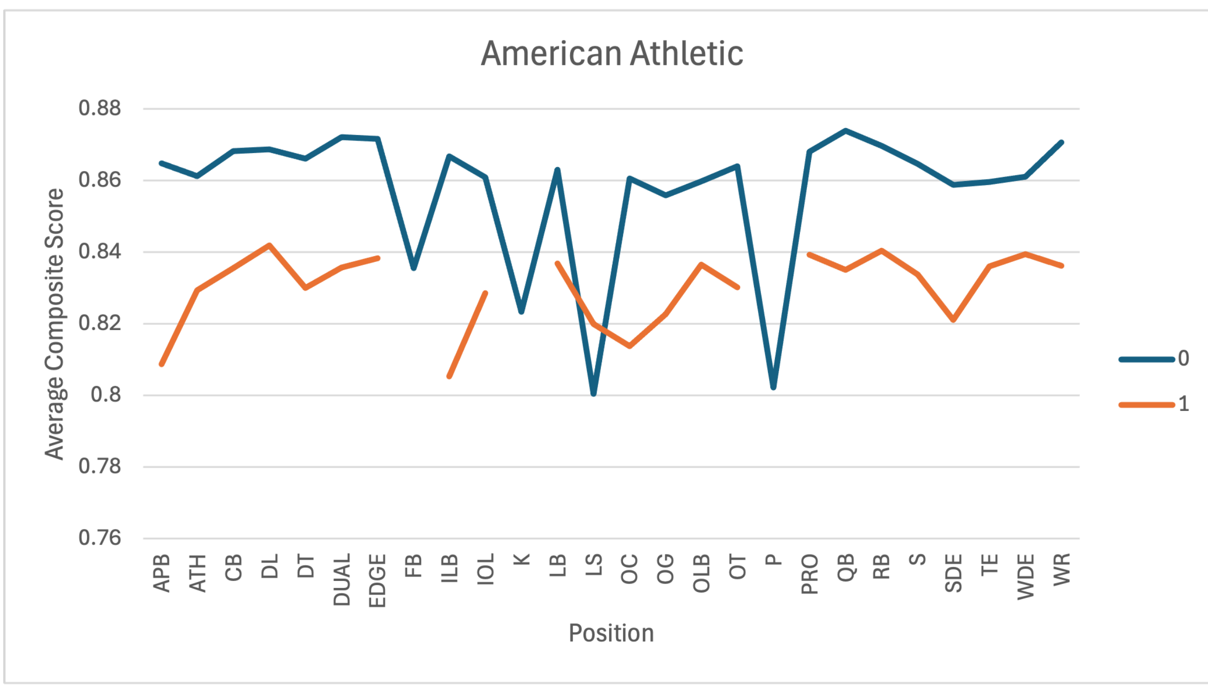

Figure 16: Average Composite Scores by Position in the American Athletic Conference

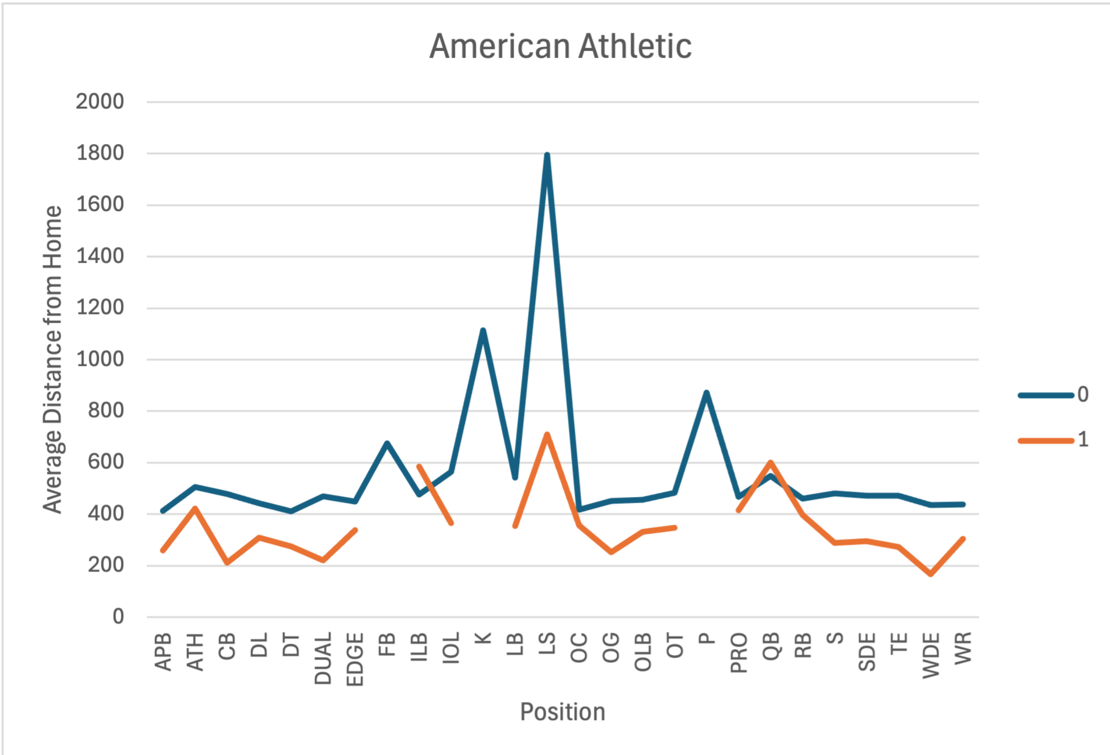

Figure 17: Average Distance From Home by Position in the American Athletic Conference

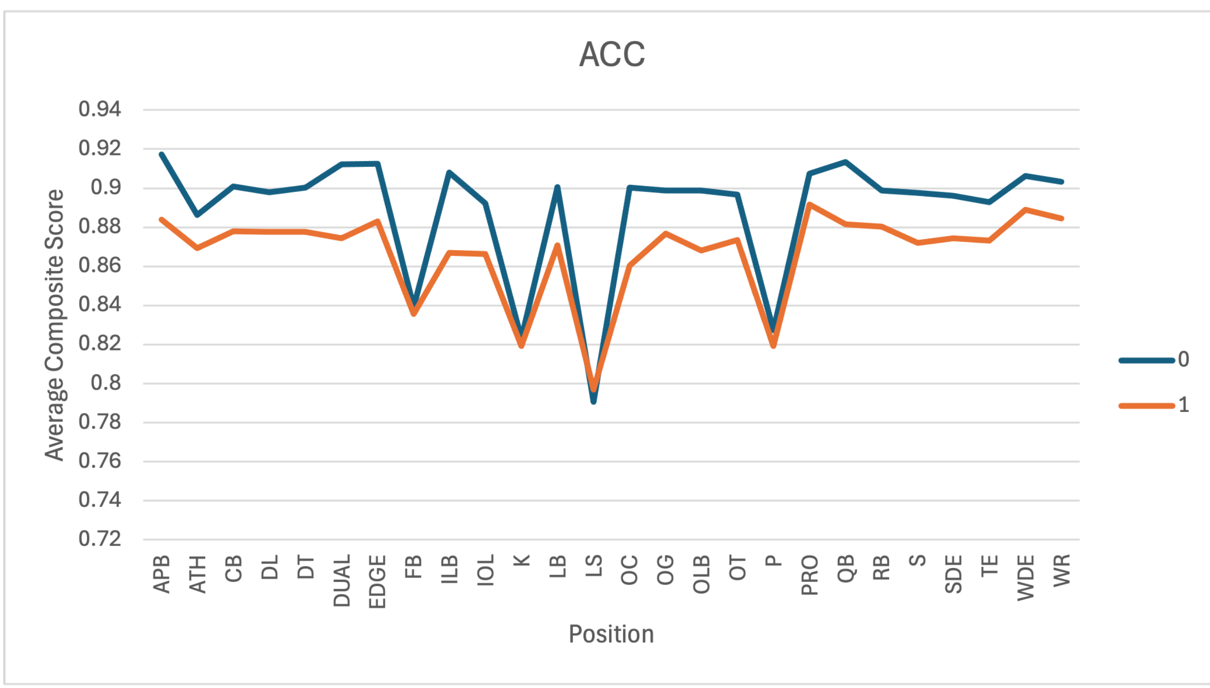

Figure 18: Average Composite Scores by Position in the American Coastal Conference

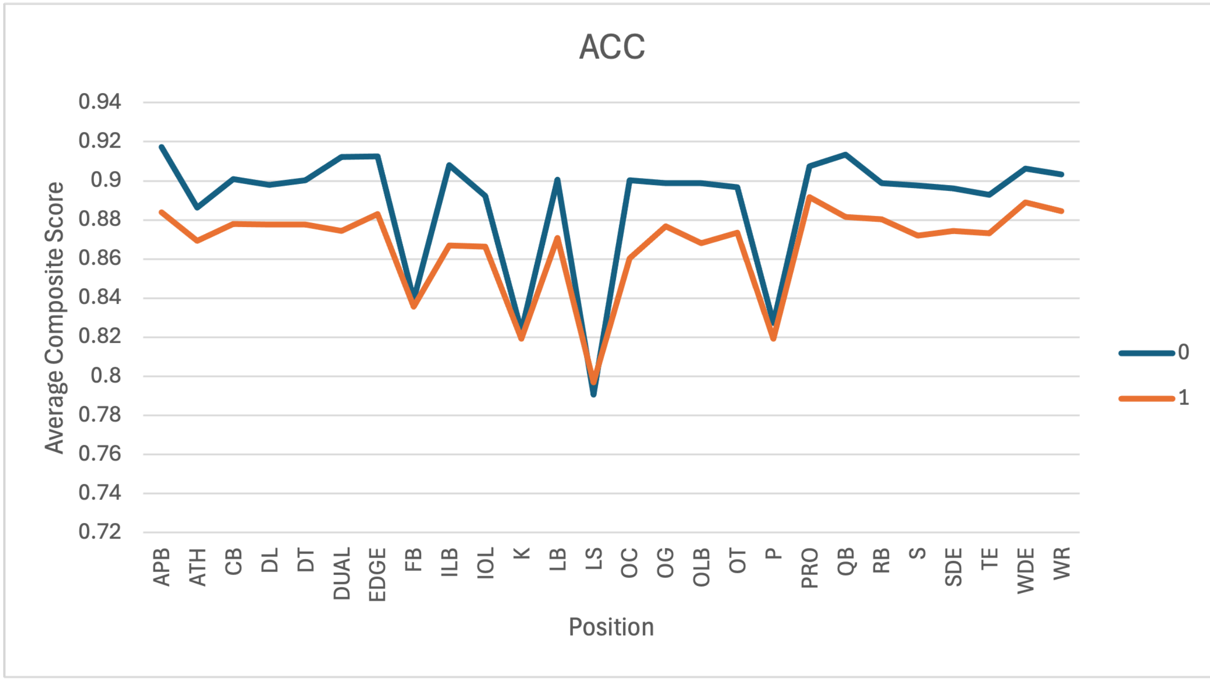

Figure 19: Average Distance From Home by Position in the American Coastal Conference

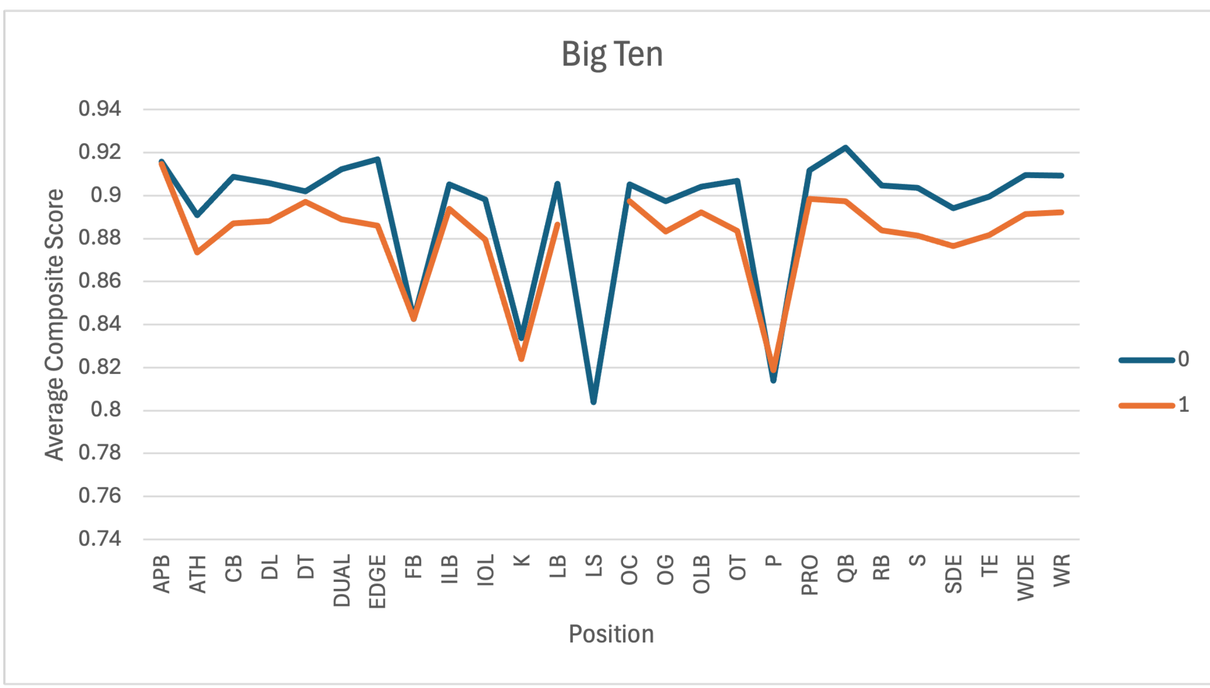

Figure 20: Average Composite Scores by Position in the Big 10 Conference

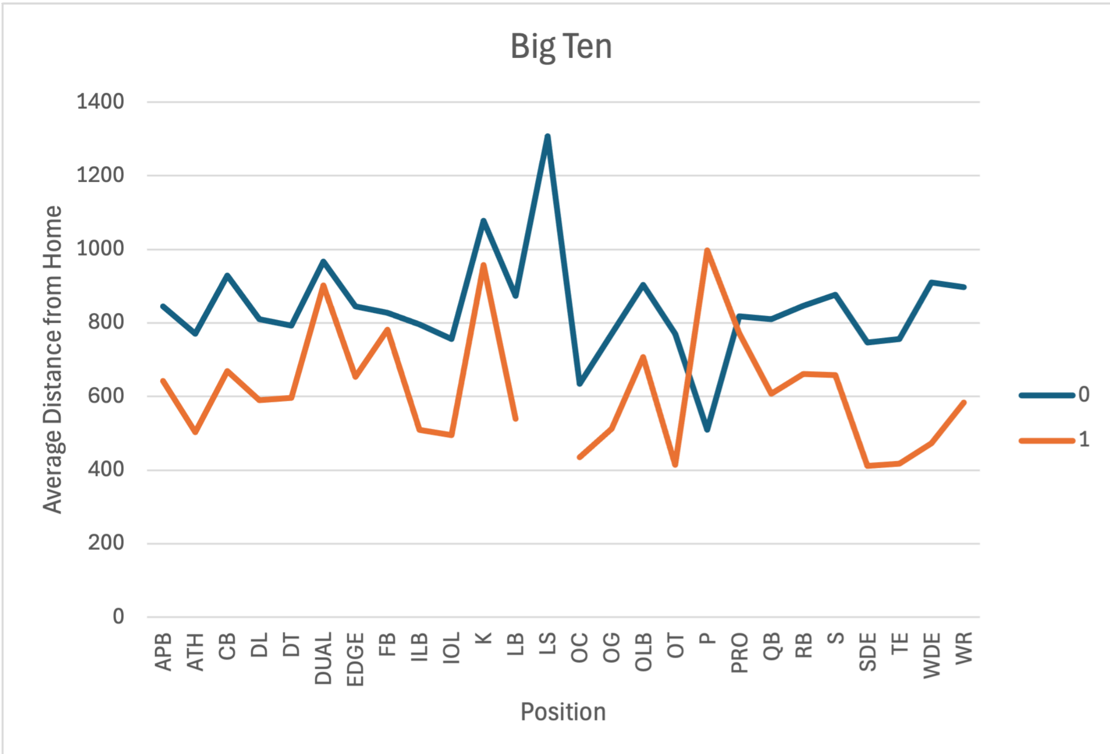

Figure 21: Average Distance From Home by Position in the Big 10 Conference

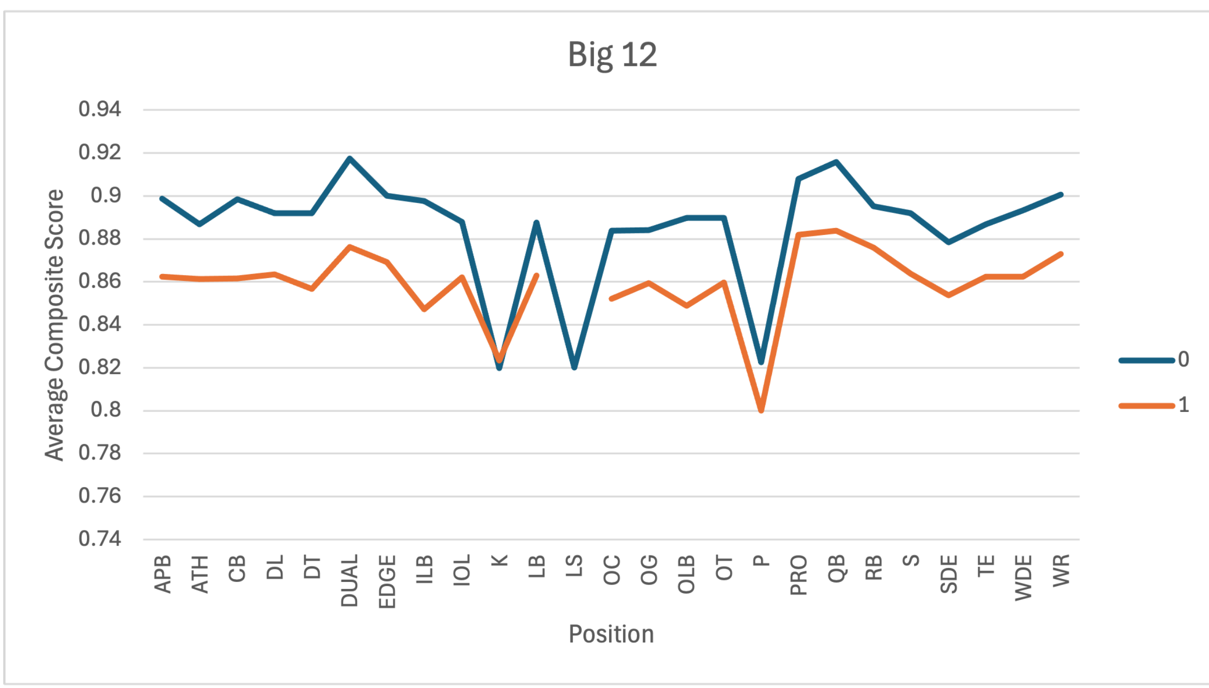

Figure 22: Average Composite Scores by Position in the Big 12 Conference

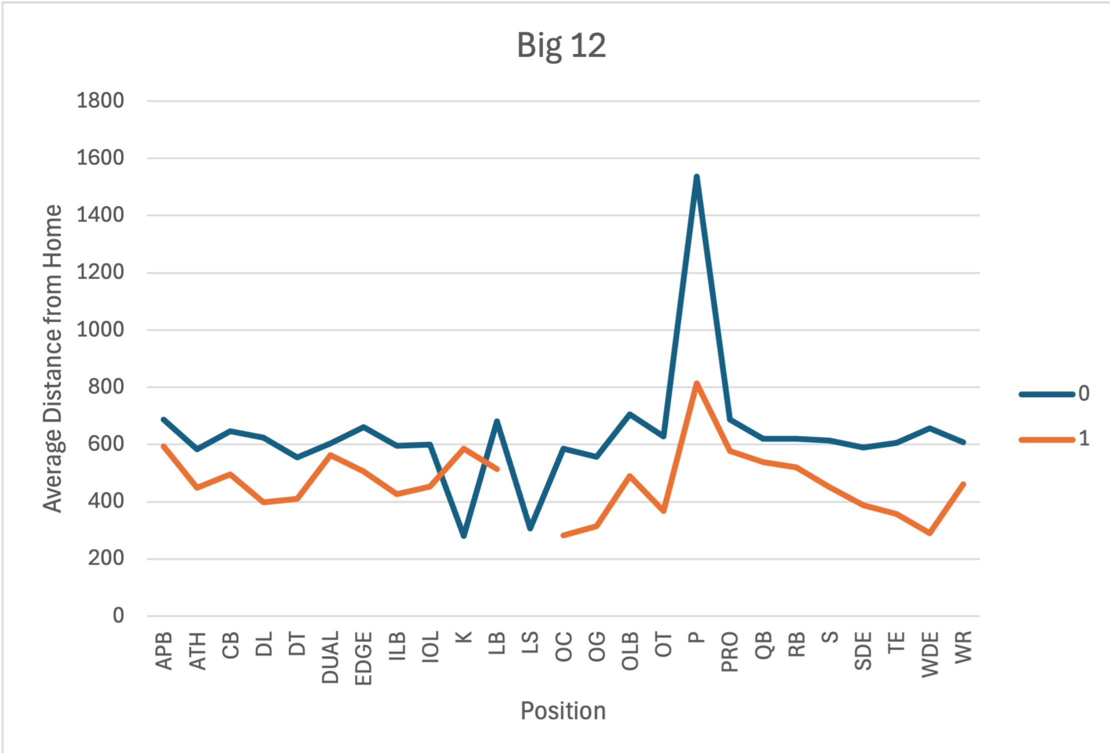

Figure 23: Average Distance From Home by Position in the Big 12 Conference

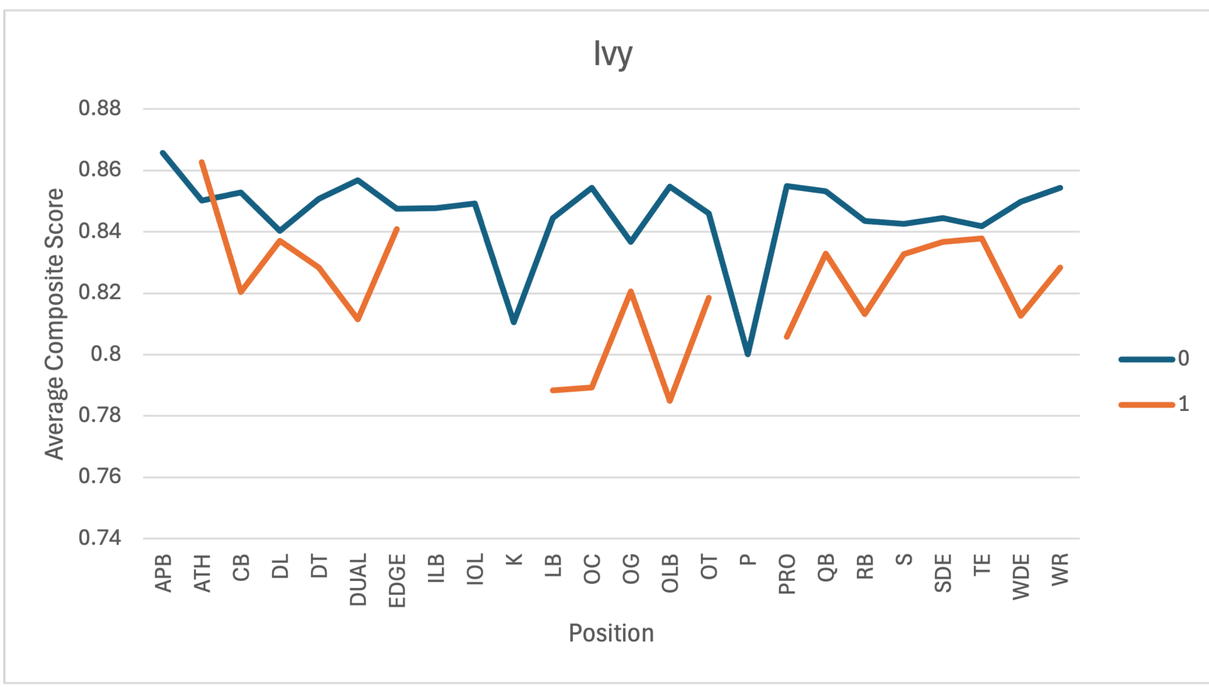

Figure 24: Average Composite Scores by Position in the Ivy League Conference

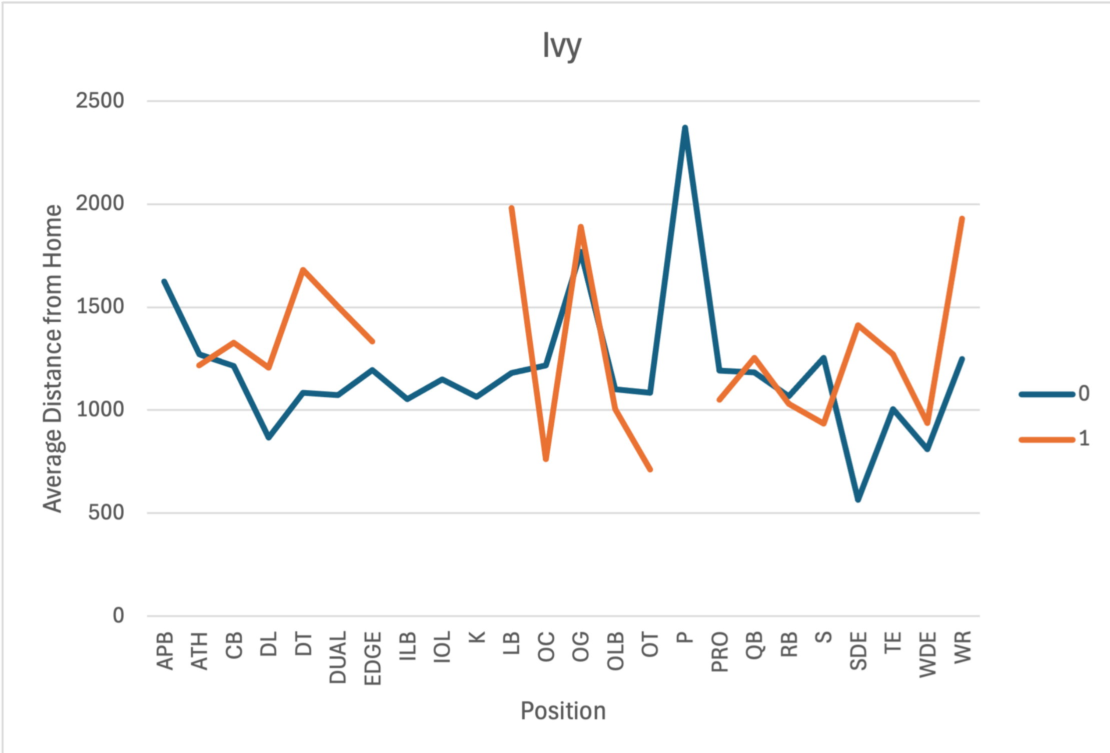

Figure 25: Average Distance From Home by Position in the Ivy League Conference

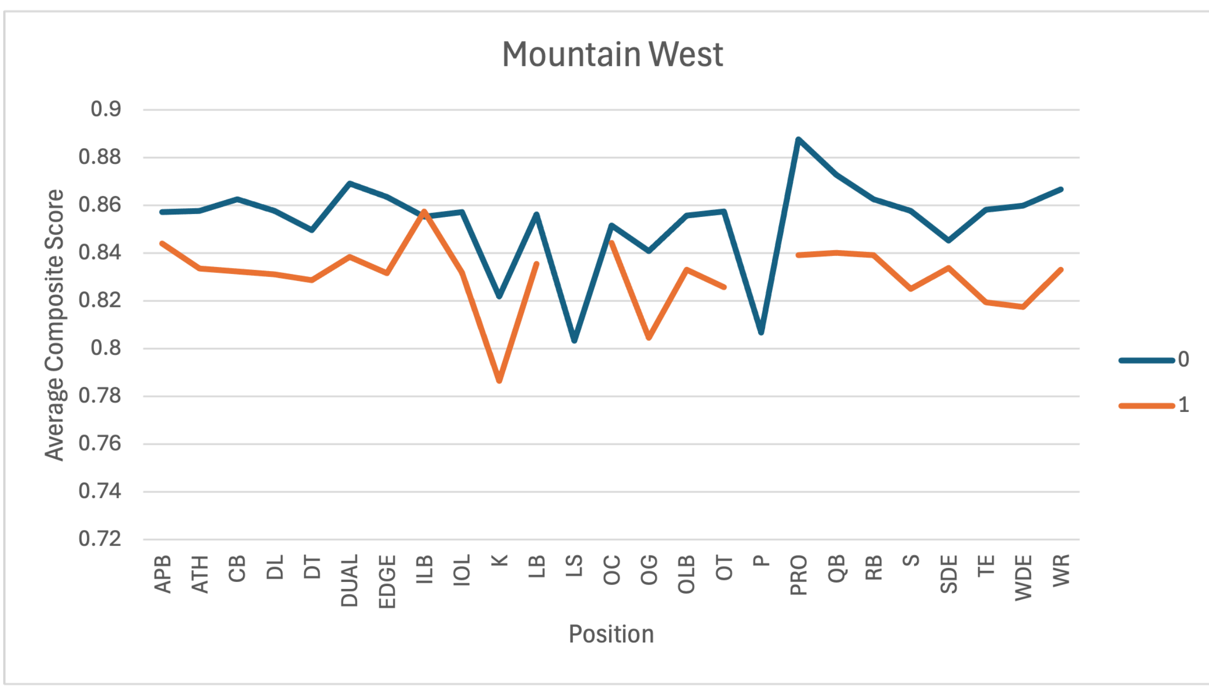

Figure 26: Average Composite Scores by Position in the Mountain West Conference

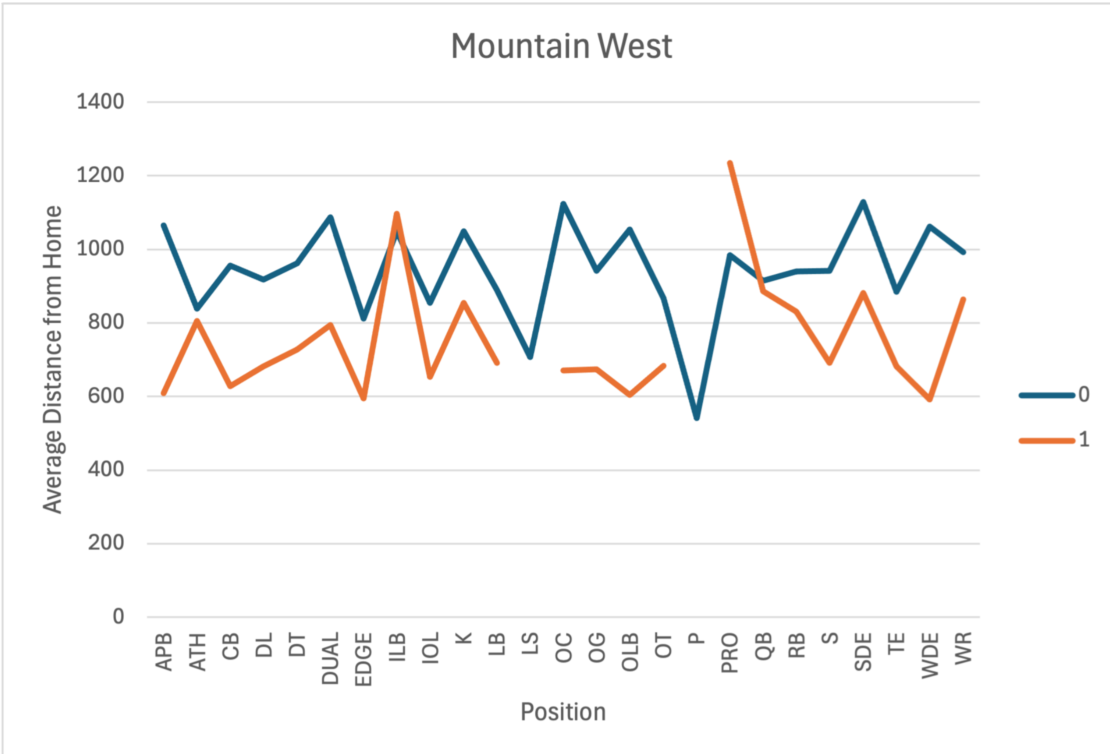

Figure 27: Average Distance From Home by Position in the Mountain West Conference

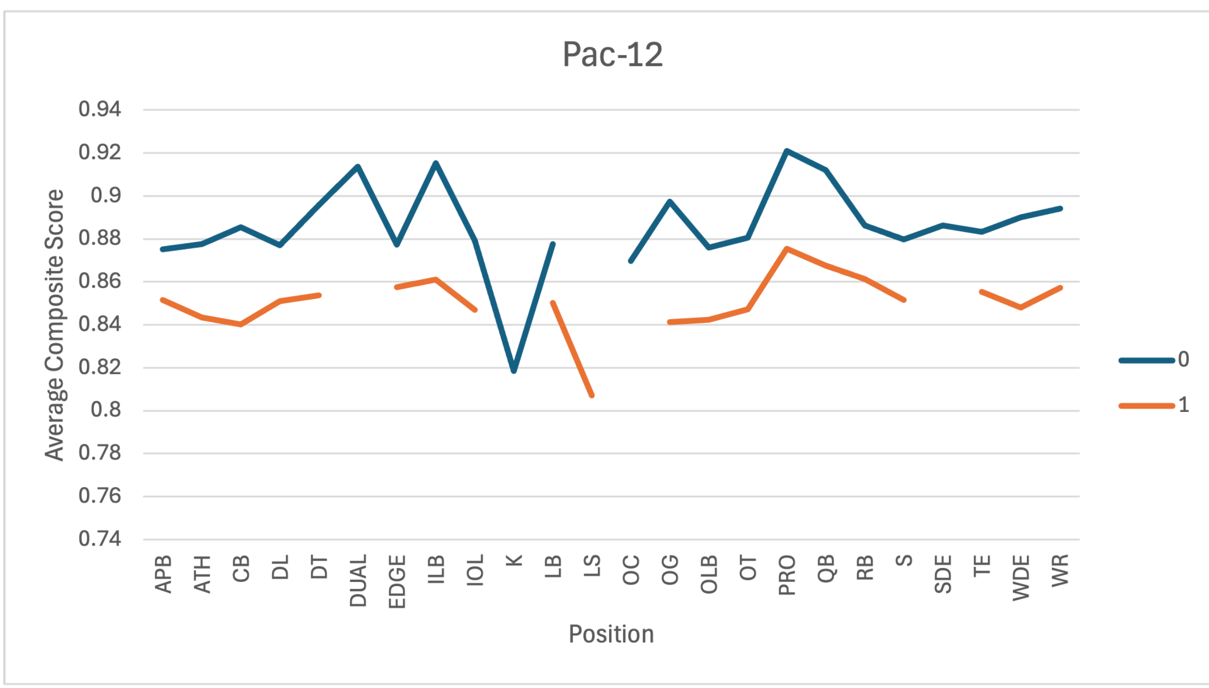

Figure 28: Average Composite Scores by Position in the PAC 12 Conference

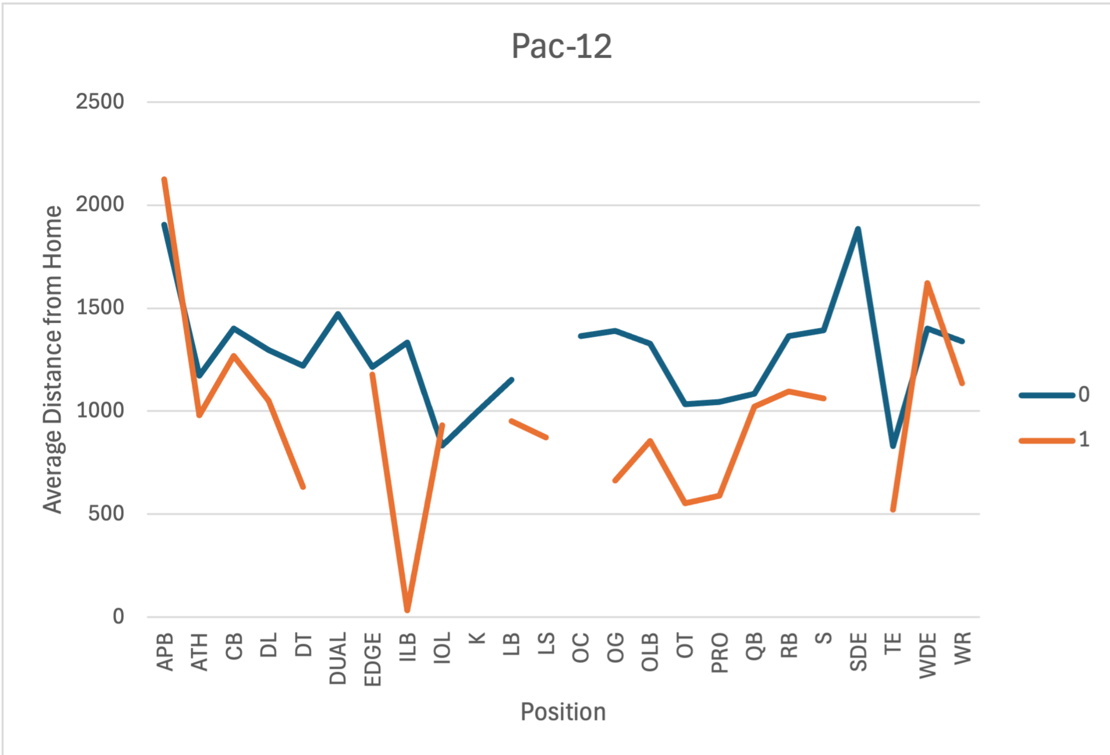

Figure 29: Average Distance From Home by Position in the PAC 12 Conference

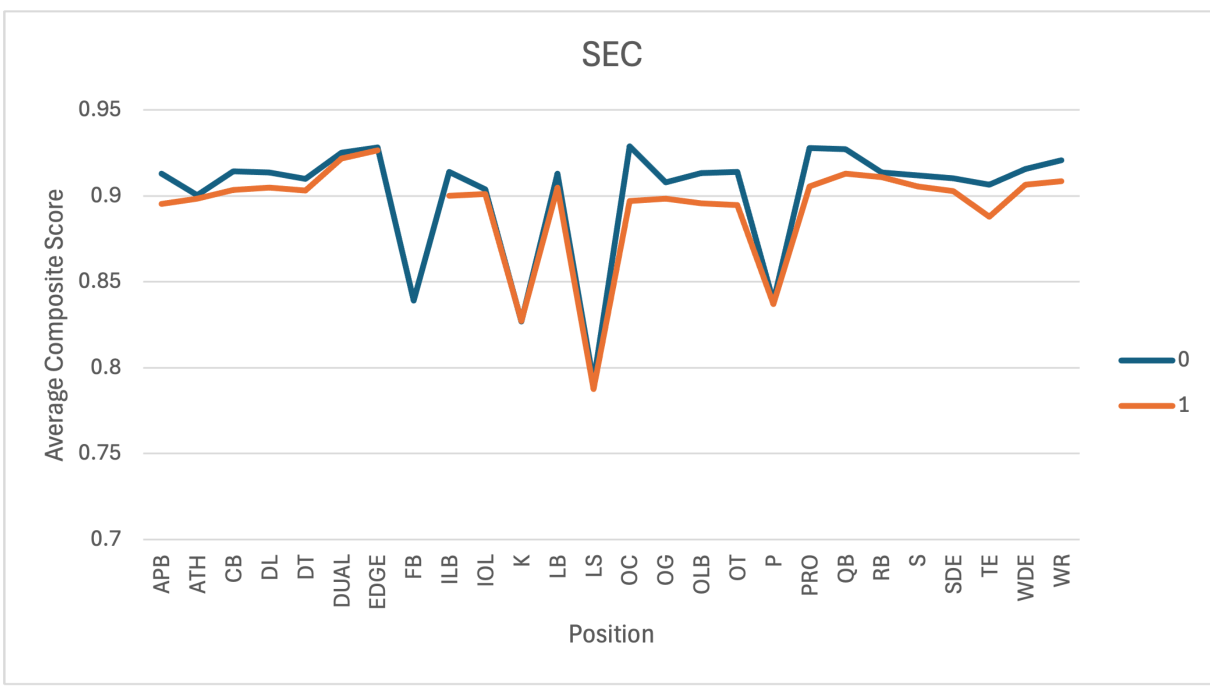

Figure 30: Average Composite Scores by Position in the Southeastern Conference

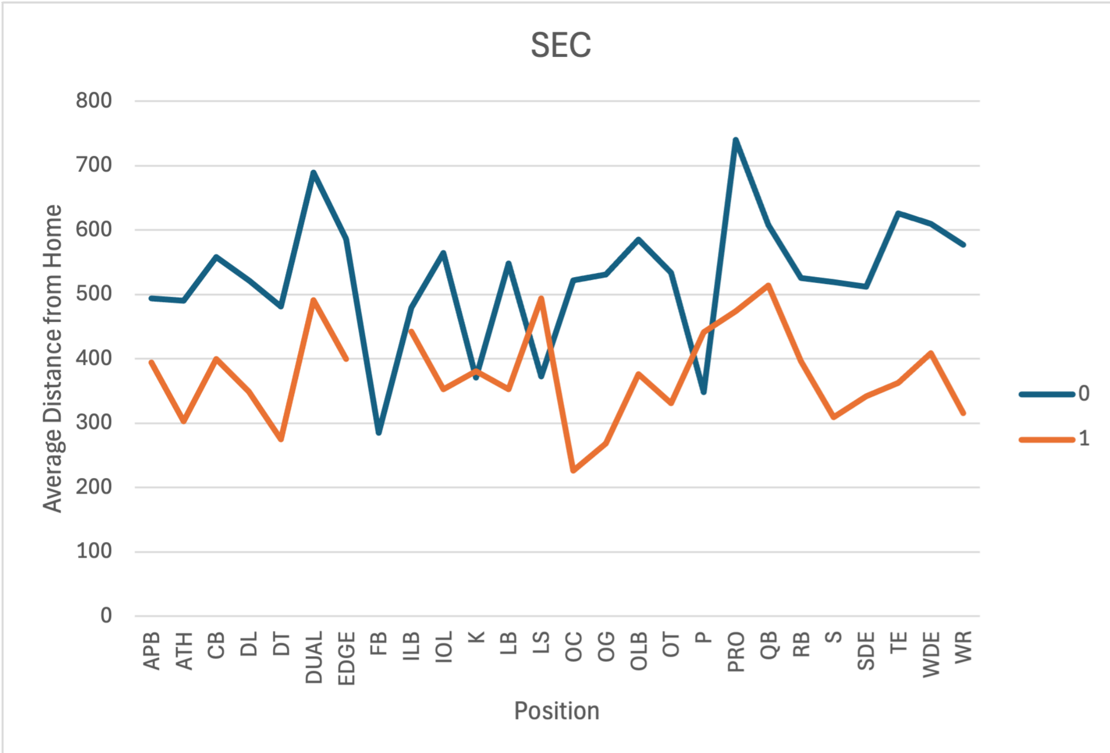

Figure 31: Average Distance From Home by Position in the Southeastern Conference