**For the Win: Understanding the Most Important Factors in NCAA Football Outcomes in the 2022 Season**

Justin Carter

University of Wisconsin – Superior

DS785: Capstone

Dr. Thomas Buck

February 18, 2023

**Abstract**

Each year, millions upon millions of college football fans flock to stadiums or televisions to watch what they consider to be the most exciting competition in American sports. While some may argue that the game appears simple to the eye, those with more intimate knowledge of the game would argue that there are hundreds, if not thousands of variables that contribute to game outcomes. A subsection of these variables is examined in order to provide a case study of the 2022 NCAA football season. Pursuant to that task, this work provides lists of the top 10 most influential team statistics on a weekly basis for the 2022 season. Additionally, this work endeavors to create a predictive random forest model of the most influential team variables throughout the course of the 2022 season. This model was able to correctly predict the outcome of more than 70% of games in the season. The results of this work may be of interest to a wide variety of individuals and organizations, especially college football programs and sports betting outlets, who may find the information regarding influential variables in game outcomes useful in adjusting their current strategies. The paper concludes by discussing the implications of these variables and placing the results of the analysis in proper context.

*Keywords:* NCAA football, 2022 season, random forest, team statistics, game outcomes

**Table of Contents**

**Introduction6**

Background of the Problem8

Significance of the Study8

Limitations10

**Literature Review11**

Introduction11

Prior Works12

Academic Work I12

Academic Work II12

Academic Work III13

Academic Work IV14

Academic Work V14

Academic Work VI15

Academic Work VII16

Academic Work VIII17

Academic Work IX18

Academic Work X18

**Methodology19**

Introduction19

Research Design19

Research Question20

Data Collection20

Data Analysis20

**Results22**

Introduction22

Results23

Week 223

Week 324

Week 424

Week 525

Week 626

Week 726

Week 827

Week 928

Week 1028

Week 1129

Week 1230

Week 1330

Bowls 131

Bowls 232

Combined Dataset and Model Results32

**Discussion**3**4**

Introduction34

Recommendations35

Conclusion36

**References37**

**Appendix43**

**List of Tables**

Table 1: Summary of Datasets Prepared for Analysis22

Table 2: Ten Most Influential Team Statistics: Combined33

**For the Win: Understanding the Most Important Factors in NCAA Football Outcomes in the 2022 Season**

**Introduction**

Although often considered a uniquely American sport, the game of American football actually can be traced to various versions of association football and rugby football. The highly vaunted game was the product of several changes added to the aforementioned games, primarily by the likes of Walter Camp, who was a coach for Yale University and is considered by many to be the “Father of American Football.” Amongst the most notable changes Camp contributed to the game were rules concerning the number of players allowed on the field at a time, down-and-distance, and the forward pass. The rule changes that Camp proposed were critical in altering what had been essentially modified rugby into something more recognizable to modern football fans as the game they love. However, Camp was not the only individual that contributed significantly to American football. A key paradigm shift to the game came in the formation of the American Intercollegiate Football Association. This association was formed by representatives from Columbia, Harvard, and Princeton, and was meant to standardize rules for the game (The Professional Football Researchers Association, n.d.). As the game gained popularity, it spread to the South and the West, and college football conferences began to rise. By the 1930’s, familiar entities such as the Big Ten and SEC had been created. During the same time period, the Pacific Coast Conference, a precursor to the Pac-12 Conference, came to the fold. Prominent bowl games, which featured teams competing against each other from different conferences, such as the Rose Bowl, were in full swing during this era as well. From 1947 until the beginning of the BCS era, key relationships were developed between conferences in relation to bowl games. For instance, in this time frame, the Big Ten and Pac-10 champions began meeting each other on the field to play the Rose Bowl. However, this system changed in the 1990s due to a myriad of reasons, including that oft-times meetings between the most prominent teams in the country could not take place due to their obligation of playing their assigned conference partner in the bowl system. In 1992, bowl partners were opened up to some degree with the Bowl Coalition agreement. This, for instance, allowed champions of the Atlantic Coast Conference, Notre Dame, and the Big East Conference to play bowls with the champions of the Big Eight, Southwest, or Southeastern conferences. In 1998, the Bowl Championship Series, or BCS, was introduced to college football. This system was developed in order to pit the top two highest-ranked teams against each other in a championship game, as well as to create three other competitive games between six highly ranked teams. These teams were chosen via a new formula called the BCS standings (Fox, 2006). As with many systems that were adopted over the course of college football history, the BCS era saw many criticisms levied by fans of the sport, not the least of which was the limited number of teams that were selected to play for a championship game. In 2014, this and many other criticisms of the BCS era were addressed by the creation of the College Football Playoff. At its inception, two pairs of two college teams were invited to compete against each other. The victors of the two semifinal games then ultimately met on the field in order to determine the national champion. Throughout the four-team playoff era, speculation grew that instead of a fundamental paradigm shift, the next era of college football would entail an expansion to the current playoff structure. Media figures like Jason Gay speculated this change would occur due to the possibility of more teams being represented in the post-season affair, as well as the expanded financial opportunities for companies adjacent to the game (Gay, 2018). Despite the four-team playoff remaining the current paradigm in which a national champion is crowned, this era is indeed coming to a close. The college football playoff is on the verge of expanding to 12 teams in the 2024 season (“College football spring meeting kicks off”, 2023). While the history of college football is littered with examples of rule changes and paradigm shifts, one thing has remained constant throughout the years: the desire to more deeply understand which factors contribute to college football game outcomes.

**Background of the Problem**

For nearly as long as sports have existed, there has been a desire to utilize various methods in order to provide a greater understanding of sports outcomes. As time has marched on, numerous techniques have been developed to accomplish this task. Areas such as data analysis and data science have provided ample opportunities to dissect areas of sports that were previously left to simple intuition. Data science and data analysis began to be applied rigorously to great effect; teams began to understand which areas of a game were paramount and how to improve in these areas in order to increase their odds of a favorable outcome. However, each new season brings with it challenges in understanding which factors contribute to game outcomes. It is often difficult to weed out factors most key in predicting sports outcomes given the sheer number of factors available for analysis.

**Significance of the Study**

**G**iven the popularity of the game of college football, the game has provided a vast amount of financial opportunities for the players and programs that compete in it, as well as several businesses that are tangential to the game. Many of these financial opportunities are tied to the large number of viewers the sport draws. With events like the 2021 NCAA football championship game drawing 22.6 million viewers (a 19% increase from the year prior), it is not difficult to imagine that financial opportunities abound (Saul, 2022). College football programs, as one might imagine, benefit greatly from these opportunities. An assessment of several college football teams’ theoretical valuations on the open market was conducted in 2016. The most valuable program according to the assessment was Ohio State at a valuation of $946.6 million. Ohio State was followed by Texas at $855 million and Michigan at $811 million respectively (Beaton, 2016). A portion of what makes these programs so valuable is the opportunity to appear in post-season events like the college football playoff and other bowl games. For instance, while every conference divides payouts for postseason appearances differently in modern times, conferences whose teams reached the college football playoff semifinal games received approximately $6 million per team in the 2022 season. Conferences whose teams played in non-playoff bowls were paid $4 million per team (“College football playoffs payouts 2022-2023”, 2023). Another avenue for financial opportunity for college football programs lies in ties with companies that run the gambit from providing nutrition to outfitting teams’ players. For instance, in 2019, Forbes listed the most lucrative college apparel deals. These deals were headed by UCLA’s contract with Under Armor for a cash and product allowance of $12.7 million (covering the years 2017-2032) and Louisville’s deal with Adidas for a cash and product allowance of $10.66 million (for the years 2018-2028) (Kleinman, 2019). In recent times, college football players have become able to take advantage of the multitude of financial opportunities that exist within the sport as well. This is a development that occurred due to a 2022 Supreme Court ruling that found that the NCAA had violated antitrust rules (Mullaney, 2021). From national fashion brands to local car dealerships, companies that were once barred from creating financial partnerships with college football players were freed to do so. These partnerships could result in significant incomes for college football players, as was speculated in an analysis conducted in 2020. The analysis estimated that the top players in NCAA football could earn as much as $2.4 million per year at the highest end (Huddleston, 2020). The game of NCAA football also presents significant opportunities for companies that specialize in sports betting. Although the scope of sports betting had been severely limited to a small number of locales in the past, this is no longer the case in the United States. After a Supreme Court ruling that allowed sports betting in states outside of Nevada in 2018, more than half of US states have since legalized it (“The sports betting market in America is exploding”, 2022). Given a broader range of opportunities, it is no wonder why brands like FanDuel continue to find success. The brand saw its first quarterly profit in the second quarter of 2022. This came despite aggressively investing in advertising and customer acquisition. Indeed, the brand expects to see full-year profitability in 2023 in spite of the possibility of a recession (Schafer, 2022). However, this growth is not just limited to FanDuel. Since 2018, $95 billion has been wagered in legal sports bets throughout the country, and by the year 2028, the sports betting market is predicted to be worth $140 billion (“Sports betting in America is exploding”, 2022). One need only look at famous works like *Moneyball* to begin to understand the opportunities that proper analyses of sports can provide. Naturally, analyses of this kind do not only provide vital information to the teams and individuals directly participating in the sport in question; institutions and businesses tangential to those sports also stand to benefit from these analyses. Sports outlets such as ESPN and Fox Sports can employ these analyses in order to display key information such as the likelihood that a team or an individual prevails in a game. Sportsbooks, too, benefit from this information via a greater ability to attribute betting lines on various outcomes.

**Limitations**

This work aims to provide an analysis of NCAA FBS football team statistics and their relative importance in determining NCAA football outcomes (win/loss) in the 2022 NCAA football season. This area of statistics was chosen for analysis in order to provide a focused scope for outcome analysis. Thus, while this work aims to provide a look into the most influential factors in NCAA football outcomes, other notable variables that may influence the game, such as individual statistics, will be left for future investigations. Relative factor importance will be recorded and assessed, and those with the highest predictive value will be utilized in order to create a statistical model to predict NCAA football outcomes.

**Literature Review**

**Introduction**

Like many topics worthy of statistical analysis, the game of football is quite complex. From team rankings and team-level statistics to individual players and intra-play factors, there are numerous aspects of the game that can present as valid options for predicting game outcomes. Naturally, given the vast amount of variables the game presents for analysis, there is no shortage of methodologies that have been developed to incorporate them. In some academic works, researchers have opted for delving into team ranking methodologies in order to predict game outcomes. Others have focused on specific team variables (such as prior win percentage and average team performance) to build prediction models. As there has been a massive effort by researchers to create predictive models of football game outcomes for several decades, an exhaustive review of each piece of literature concerning this area of research would likely be infeasible. Instead, a collection of prior research concerning the prediction of football game outcomes, involving a variety of methodologies, will be presented in chronological order to properly situate and contextualize this work within the broader academic conversation regarding this topic.

**Prior Works**

***Academic Work I***

With their academic contribution in 1978, Vergin and Scriabin aimed to develop profitable strategies for betting on National Football League games. While other works in predicting football games have utilized variables like teams’ relative strengths and weaknesses in certain areas compared to others, the authors instead endeavored to examine how point spreads for football games were made. They noted that point spreads for games are in part a reflection of predicted game outcomes. Like other areas of prediction, the authors argued, errors in those predictions are not always random, and can at times have an underlying pattern due to prediction biases. As point spreads are made by various companies for the purpose of sports betting, biases in point spreads could range from unconscious favoritism of certain teams to the conscious financial consideration of the companies offering those spreads. The researchers utilized data from the years 1969-1972 for detecting possible bias patterns, along with data from years 1973-1974, in order to test the persistence of any patterns found. After conducting their analysis, the researchers found noticeable patterns of bias in point spreads. In doing so, they developed a number of proposed betting strategies, including the recommendation to bet on underdog teams when the point spread is above 5 points (Vergin & Scriabin, 1978).

***Academic Work II***

In 1980, Stefani contributed to the growing work concerning the prediction of sports outcomes via his paper “Improved Least Squares Football, Basketball, and Soccer Predictions.” As part of this research, Stefani surveyed sports rating systems, including his own least squares system. Utilizing said system, he was able to make game outcome predictions via the rating difference of future opponents. The system Stefani proposed was implemented in order to simplify a previous least squares system he made years prior. This system was applied to nearly 9000 games. These games ranged in domain from college basketball, to community college football, as well as the World Cup of Soccer. Notably, the work aimed to improve upon his previous system by including a factor that reduced predicted win margin as compared to rating difference, and by including consideration for home team advantage. With this system, Stefani was able to make correct outcome predictions for about 70% of the nearly 9000 games to which the system was applied (Stefani, 1980).

***Academic Work III***

In this work, Stern focused on elements of NFL football games that had the potential to influence game outcomes. The author begins by noting that several prior works including those from Pankoff, Vergin & Scriabin, and Zuber, Gandar, & Bowers have examined the statistical relationship between National Football League game outcomes and their assigned point spread values. These analyses, in general, found that significant results incorporating these values were difficult to obtain given that football games tend to have high variance (Pankoff, 1980; Vergin & Scriabin, 1978; Zuber, Gandar, & Bowers, 1985). The author aimed to discuss methods that could model NFL game outcomes as a normal random variable with its mean equal to the point spread. For his analysis, Stern utilized a data set of the scores and point spreads of each NFL game during the 1981, 1983, and 1984 seasons. The data set contained 672 National Football League games. For this analysis, Stern framed his research question as “What is the probability that a team favored to win a football game by p points wins the game?” Stern determined that the margin of victory by a favored team can be approximated as a Gaussian random variable with a mean equal to the point spread and standard deviation estimated at 13.86. After this primary analysis, Stern confirmed the validity of his findings utilizing later data from the 1985 and 1986 seasons (Stern, 1991).

***Academic Work IV***

Fair and Oster added to this body of work with their paper “College Football Rankings and Market Efficiency.”The general goal of this paper was to apply optimal weights to various existing college football ranking systems. These weights would be used in order to make a predictive system that was superior in predictive ability to any of the individual ranking systems included in their analysis.The authors of this paper compared 9 algorithm-based team ranking systems, including Sagarin’s USA Today rankings, the Seattle Times/Anderson & Hester rankings, as well as the Atlanta Journal-Constitution Colley Matrix system rankings.Data were collected for the years 1998, 1999, 2000, and 2001, utilizing 10 weeks of outcome data starting with week 6 of each season. The data included in their work was limited to the 117 Division I-A teams that existed at that time, and 1588 games in total. After excluding non-common observations between the systems, the team observed very high correlation coefficients between the chosen ranking systems, ranging from .779 to .973.Following the assignment of appropriate weights to the various ranking systems chosen for this work, the authors were able to create a model that predicted college football outcomes with a 74.7% accuracy (Fair & Oster, 2002).

***Academic Work V***

West & Lamsal’s work aimed to apply linear modeling techniques for the prediction of college football bowl outcomes. In particular, the paper aimed to consider whether a modeling approach incorporating team-level information would be able to accurately predict future outcomes of college football games. The authors collected data from a number of publicly available sources (NCAA, cfbstats.com, etc) before the onset of the bowl games. Team statistics were collected concerning teams who were invited to play in bowl games in seasons 2004-2007. The collected statistics included:

1. Number of games played
2. Scoring margin
3. Offensive yardage accumulated per game
4. Offensive first downs per game
5. Defensive yardage yielded per game
6. Defensive first downs yielded per game
7. Defensive touchdowns yielded per game
8. Turnover margin
9. Strength of schedule (computed by Jeff Sagarin for USA Today)

Values of the above variables were standardized within each year for teams selected to compete in bowl games. The linear model incorporating the aforementioned team statistics yielded an overall outcome of 59.4%, accurately predicting 19 of 32 bowl games (West & Lamsal, 2008).

***Academic Work VI***

In another piece concerning the prediction of football outcomes, LoPilato et al. focused on the variance of performance of college football teams (peak, typical, and variability) in predicting win percentage, home game attendance, and bowl game payouts. For the purposes of the study, LoPilato et al. utilized specified definitions of the terms “peak”, “typical”, and “variability.” They defined typical performance as the average level of performance over a given number of performance episodes. Peak performance was operationalized as the highest team statistical performance in any given game over the course of a season. Finally, they operationalized performance variability as the difference in performance between games in a season. LoPilato et al. performed their work via data obtained from an archival sports database that included 193 Bowl Subdivision teams from three separate seasons. While all types of performance were considered (peak, typical, and variable) together, they found that only typical team performance predicted win percentage, and explained 49% of the variance within the data collected. In addition to this, it was found that both typical and peak team performance were associated with win percentage, bowl game payout, and attendance. Team offensive variability was associated with win percentage, but not with fan attendance or bowl game payout. Team defensive variability was found to be unrelated to any of the three outcomes the authors were measuring (LoPilato et al., 2013).

***Academic Work VII***

The work conducted by Leung & Joseph also aimed to create a model for predicting college football outcomes. The authors comment on the various methods of approaching the prediction of game outcomes and note that many of the analyses available include subjective and anecdotal information and emphases. The analysis they conduct, much like this work, aims to take a more quantitative approach to game predictions. The authors also note that making direct predictions between teams in college football can be difficult due to varying player availability as well as the fact that most teams only play each other once in a given season. The authors devised a data mining approach wherein they analyzed a set of teams that were most similar to each of two competing teams, found the results of games between the teams in those sets, and utilized those results to predict the outcome of the game between the original two teams. The authors’ data mining approach to predicting college football games led to an accuracy of 91.43% for the 2010-2011 NCAA football season. The accuracy was compared to a number of other models for the 2010-2011 season, including those of Whatifsports.com (57.14%), a neural network approach (71.43%), and a decision tree approach (82.86%) (Leung & Joseph, 2014).

***Academic Work VIII***

Of course, there are a variety of ways to predict football game outcomes.Yurko et al., in their work, set out to predict game outcomes in a relatively novel manner; their work focused on the prediction of game outcomes via a continuous-time within-play methodology.As noted in their research, discrete-time estimates of the value of plays are a more common method of outcome prediction due to most publicly available football data being recorded on a play-by-play basis.The authors note that while statistics like win probability and expected points may be useful, there has been no prior research into how those values change within plays. Yurko et al. were able to take advantage of the relatively recent advent of play-tracking technology in order to measure these variables.In particular, their paper makes two contributions. The first contribution was a general framework for within-play valuation of game outcomes in NFL football. The second contribution made was a model which predicts the yards gained from a ball carrier given their position, as well as the positions and trajectories of their teammates and the other team’s players.For their analysis, the team utilized a data set containing NFL tracking data from the first 6 weeks of the 2017 season.Their analysis included training XGBoost and LASSO models on their entire dataset. Both models determined that the most important factors included in the dataset were the ball carrier’s distance to the closest defensive player, as well as the current speed of the ball carrier (Yurko et al., 2020).

***Academic Work IX***

In 2020, researcher Rick Wilson contributed to the academic pursuit of predicting college football games via his paper “College Football Overtime Outcomes: Implications for In-Game Decision-Making.”As implied by the title, this work focused on predicting and recommending strategies for college football teams that find themselves competing in overtime games.For his analysis, Wilson opted to utilize both regression and decision tree algorithms. Wilson’s analysis utilized SAS’s default parameters in order to generate the logistic regression model. The logistic regression model employed ridging for optimization. His decision tree model employed the default parameters found in Enterprise Miner, including the use of entropy criteria for ordinal splits. According to Wilson, these parameters were chosen in order to aid in the ability to replicate his results.Seven years of college football game data were utilized for the analysis, with 243 games included therein.The logistic model employed to predict game outcomes correctly predicted 59.6% of the games included in the analysis. The decision tree approach the author employed achieved a classification accuracy of 66% (Wilson, 2020).

***Academic Work X***

Naturally, there are several different techniques and methodologies with which one can predict the success or failure of a given team. For their part, Mankin et al. performed an analysis focusing on the role that college football recruiting rankings played in team success.The team compiled 247 composite rankings from recruiting classes ranging from 2002-2018. These rankings would be juxtaposed to Sangarin team rakings ranging from the 2005-2018 college football seasons. The difference in time ranges between recruiting rankings and team performance was put in place due to the average college football player remaining with their respective team for about 4 years.After conducting their analysis, it was found that up to 36% of the variation in the Sagarin ratings was explained by recruiting rankings (Mankin et al., 2021).

**Methodology**

**Introduction**

As previously mentioned, the purpose of this work was to perform an analysis of NCAA FBS teams for the 2022 Season. This section intends to delve into the research design, research questions, data collection process, and data analysis process undertaken pursuant to this endeavor.

**Research Design**

Team statistics from the week prior were complemented with game outcomes in a cascading fashion in order to properly conduct the analysis. For instance, for the Georgia vs. Ohio State bowl game in the 2022 season, Ohio State’s team statistics, as well as Georgia’s team statistics up until that game were put in a row of a dataset, along with the winner (1 for a home team win, or 0 for a home team loss).For the purposes of this analysis, team statistics are denoted by home teams and away teams. A home team is a team that “hosts” the other team in a football game. This paradigm was utilized throughout the dataset and holds true for special circumstances, such as bowl games. Again utilizing the Georgia vs. Ohio State bowl game as an example, even though the game was played outside Georgia’s home field, Georgia is still considered the home team for that game in this analysis. Initially, all games in which FBS teams were involved were considered for analysis. However, games that involved FCS teams matching up with FBS teams were eliminated from the analysis. Ultimately, 17 datasets were created from this process. They include team statistics data from week 2 of the 2022 season up until the national championship game. It was ultimately decided that week 1 of the 2022 season would be excluded from analytical consideration in order to increase the overall number of team statistics variables available for analysis. Thus, game outcome data were collected from week 3 onward, including the outcome data for the national championship game.

**Research Question**

Naturally, the research question this work intends to answer is which team statistics have the highest impact on NCAA football outcomes. Since there is a large number of metrics officially tracked by the NCAA, without proper analysis, it would be difficult to determine which of the several metrics tracked by the organization most contribute to game outcomes. An additional question this work aims to answer is how accurately these metrics can predict game outcomes for the season.

**Data Collection**

Game outcome data were collected from CBSSports.com; this data was utilized in order to provide information concerning which team in each game in the analysis won.Team statistics were collected from the NCAA’s website. Python was utilized in order to scrape game outcomes from CBSSports.com. Team statistics data was downloaded from the NCAA’s sports statistics website. R was utilized in order to create and prepare datasets for analysis, as well as to conduct the analysis.

**Data Analysis**

In order to determine the team statistics most relevant in determining game outcomes in the 2022 season, a random forest model was created for each week in the analysis, and the top 10 variables contributing to each model were recorded.The random forests technique was chosen due to the method’s ability to produce reasonable results given datasets with relatively few observations and a high number of features. Other machine learning techniques, including clustering, were initially considered for the analysis. However, given the sheer number of features to be examined, it was decided that clustering would not be a feasible avenue for analysis. With this in mind, a random forest model was also applied to a combined dataset consisting of all game outcomes and team statistics collected in order to produce an overall list of the 10 most important team statistics that determined game outcomes throughout the season, as well as to produce a model that featured higher predictive power.Table 1, which summarizes the contents of each dataset employed in this analysis can be found below. It should be noted that datasets prepared from raw game outcome and team statistical data are named for which week of team statistics are present within them. Thus, the Week2 dataset contains team statistical data from the second week of the season along with game outcomes from week 3 of the season.

**Table 1**

*Summary of Datasets Prepared for Analysis*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Dataset Name | Week of Team Statistics | Week of Game Outcomes | Number of Observations (Games) | Team Statistical Variables |
| Week2 | 2 | 3 | 45 | 379 |
| Week3 | 3 | 4 | 55 | 379 |
| Week4 | 4 | 5 | 50 | 379 |
| Week5 | 5 | 6 | 52 | 379 |
| Week6 | 6 | 7 | 46 | 379 |
| Week7 | 7 | 8 | 51 | 379 |
| Week8 | 8 | 9 | 46 | 379 |
| Week9 | 9 | 10 | 59 | 379 |
| Week10 | 10 | 11 | 62 | 379 |
| Week11 | 11 | 12 | 59 | 379 |
| Week12 | 12 | 13 | 62 | 379 |
| Week13 | 13 | 14 | 11 | 379 |
| Week14 | 14 | 15 | 1 | 379 |
| Bowls1 | 15 | 16 | 8 | 379 |
| Bowls2 | 15 | 17 | 32 | 379 |
| Week18 | 17 | 18 | 1 | 379 |
| Combined | All | All | 631 | 379 |

**Results**

**Introduction**

As mentioned previously, the objective of this work was to conduct an analysis of the 2022 NCAA football season in search of the most influential team factors on game outcomes. To that end, several weeks of the 2022 season were analyzed in order to determine these factors. Several random forest models were created for the purpose of determining these factors on a weekly basis, with the final model containing information on the Combined dataset. The 10 most important factors contributing to NCAA football outcomes were recorded, along with an accuracy measure on the final, predictive model. The results of this analysis can be viewed below on the next several pages. It should be noted that lists were not created for the Week14 and Week18 datasets, as they did not contain enough information to be analyzed individually. However, the information contained in those datasets was present in the Combined Dataset.

**Results**

***Week 2***

As can be seen below, numerous defensive team statistics comprised the list of most influential factors in game outcomes for Week 2. The home team’s yards per game allowed, their total defense rank, and the away team’s yards per play allowed were deemed particularly important:

1. Yards per game allowed by home team’s defense
2. Total defense rank of home team
3. Yards per play allowed by away team
4. Total yards allowed by home team’s defense
5. Number of blocked punts allowed by home team
6. Scoring defense rank of home team
7. Completion percentage rank of home team
8. Penalty yards per game rank of home team
9. Points after touchdown of home team
10. Points allowed by home team’s defense

***Week 3***

Per the list below, the trend of team defensive statistics comprising the most influential factors in game outcomes continued. Interestingly, all of the most influential variables were team statistics from away teams. Away teams’ rush yards allowed per game, rushing defense rank, and their number of first downs allowed were particularly influential on game outcomes.

1. Rush yards allowed per game by away team
2. Rushing defense rank of away team
3. Number of first downs allowed by the away team
4. Total yards allowed by the away team
5. Rush yards allowed by the away team
6. Away team’s rank concerning first downs allowed
7. Points allowed by away team
8. Points scored per game by the away team
9. Touchdowns allowed by away team’s opponent’s offense
10. Yards per rush allowed by the away team

***Week 4***

In week 4, the trend of team defensive statistics dominating the list of most influential factors on game outcomes came to a halt. Instead, per the list below, teams’ 3rd down conversion percentage rank, penalty yards per game rank, and number of offense red zone passing touchdowns were most influential. In general, teams’ performance regarding penalties was also highly influential.

1. Home team’s rank in 3rd down percentage
2. Home team’s rank in fewest penalty yards per game
3. Number of touchdowns made by the home team in the red zone
4. Home team’s rank in penalty yards
5. Number of penalized yards of the home team
6. Away team’s completion percentage rank
7. Away team’s turnover margin
8. Away team’s punt returns rank
9. Home team’s 3rd down conversion percentage
10. Home team’s yards per play

***Week 5***

As can be seen in the list below, home teams’ punt return yards, their opponents’ passing efficiency against them, and their passing efficiency defense rank were the most influential on game outcomes. Team statistics for home teams comprised almost all of those most influential for this week.

1. Home team’s punt return yards
2. Pass efficiency against the home team
3. Home team’s passing efficiency defense rank
4. Home team’s punt return rank
5. Yards per passing attempt against the home team
6. Home team’s passing yards per attempt
7. Home team’s passing yards allowed rank
8. Passing yards allowed per game of the home team
9. Home team’s number of 4th down attempts
10. Number of passes made against the away team

***Week 6***

For week 6, home teams’ attempted tackles for loss was the most influential team statistic when predicting game outcomes. Several zed zone team statistics also were determined to be highly influential:

1. Home team’s opponents’ attempted tackles for loss
2. Home team’s points per game
3. Home team’s rank in conversion percentage
4. Away team’s total offensive plays
5. Scores in the red zone of the opponents of the home team
6. Home team’s 4th down conversion percentage
7. Home team’s opponents’ red zone attempts
8. Home team’s opponents’ red zone field goals made
9. Home team’s opponents’ tackle yards for loss
10. Away team’s opponents’ tackle yards for loss

***Week 7***

Week 7 saw home teams’ ranks in sacks allowed as the most influential factor in game outcomes. Several team statistics involving rushing offense and rushing defense were highly influential as well:

1. Home team’s rank in sacks allowed
2. Home team’s number of rushes
3. Yards per rush allowed by the away team
4. Home team opponents’ rushing touchdowns
5. Away team’s passing yards per game
6. Away team’s 3rd down conversion percentage rank
7. Home team’s average sacks allowed
8. Home team’s passing offense rank
9. Away team’s completed passes allowed by opponents
10. Away team’s 3rd down conversion percentage allowed

***Week 8***

For games in week 8, home teams’ rushing offensive rank proved to be the most influential factor in game outcomes. A number of other rushing team statistics, as well as home teams’ sack yards allowed via opponents were also deemed important:

1. Home team’s rushing offense rank
2. Home team’s sack yards allowed via opponents
3. Home team’s rush yards
4. Home team’s rushing yards per game
5. Away team’s number of 3rd down conversions
6. Away team’s first downs offense rank
7. Away team’s yards per rush
8. Away team’s 3rd down conversion percentage rank
9. Away team’s rushing first downs
10. Away team’s total offensive yards

***Week 9***

Team statistics concerning punting were highly present in the most influential factors in outcomes. Particularly, away teams’ opponents’ punt return yards was the most influential:

1. Away team’s opponents’ punt return yards
2. Away team’s opponents’ 3rd down attempts
3. Completion percentage rank of home team
4. Home team’s punt yards
5. Away team’s scoring defense rank
6. Away team’s punt return defense rank
7. Number of punts by the home team
8. Home team’s completion percentage
9. Offensive touchdowns allowed by the away team
10. Away team’s average score allowed

***Week 10***

The most influential factors in predicting game outcomes were quite varied for week 10. The number of away teams’ rushing first downs, home teams’ net punting rank, and away teams’ overall number of first downs comprised the top 3. Other team statistics, particularly involving rushing, were influential as well:

1. Away team’s rushing first downs
2. Home team’s net punting rank
3. Away team’s number of first downs
4. Away team’s rushing yards per game
5. Away team offense’s red zone attempts
6. Home team’s net punting yards
7. Away team’s rushing offense rank
8. Away team’s sacks allowed rank
9. Home team’s offensive yards per game
10. Away team’s first down offense rank

***Week 11***

In week 11, rushing team statistics dominated the list. The most influential factor was determined to be the number of opponent rushing touchdowns against away teams:

1. Away team’s opponents’ rushing touchdowns
2. Away team’s rushing defense rank
3. Rush yards allowed per game by away team
4. Away team’s tackles for loss
5. Rush yards allowed by the away team
6. Number of rushes made by the away teams’ opponents
7. Rushing first downs made by the away teams’ opponents
8. Away team’s tackles for loss per game
9. Away team’s opponents’ red zone rushing touchdowns
10. Away team’s team tackles for loss rank

***Week 12***

The most important factor in predicting week 12 game outcomes was a home team’s completion percentage, with the second most important factor being a home team’s completion percentage rank. Interestingly, an away team’s number of extra points also made the list.

1. Home team’s completion percentage
2. Completion percentage rank of home team
3. Away team’s kickoff return defense rank
4. Number of home team’s opponents’ punt returns
5. Number of rushing touchdowns by the home team
6. Number of 3rd down attempts by the home team
7. Red zone rushing touchdowns by home team’s opponents
8. Red zone passing touchdowns by home teams’ opponents
9. Away team’s opponents’ extra points
10. Away team’s average kickoff yards allowed

***Week 13***

Week 13, which contains several conference championship games, saw some rather interesting team statistics become prominent in influencing outcomes. An away team’s opponents’ number of punt returns, as well as a home team’s opponents’ number of kickoff returns, topped the list. Notably, third in the list was an away team’s fewest penalty yards per game rank.

1. Away team’s opponents’ punt returns
2. Home team’s opponents’ kickoff returns
3. Away team’s fewest penalty yards per game rank
4. Points allowed by away team
5. Away team’s fewest penalty yards rank
6. Away team’s solo tackles for loss
7. Away team’s average score allowed
8. Away team’s penalty yards per game
9. Yards per play allowed by away team
10. Away team’s passing yards allowed per attempt

***Bowls 1***

For the first collection of bowl games, the first two most influential team statistics were quite familiar, albeit slight variations from week 13. Notably, home teams’ red zone offensive rank and red zone scoring percentage statistics were significant in game outcomes.

1. Away team’s opponents’ punt return yards
2. Away team’s opponents’ kickoff return yards
3. Home team’s total offense rank
4. Home team’s red zone scoring percentage
5. Away team’s red zone offense rank
6. Number of touchdowns scored by the home team
7. Away team’s rushing offense rank
8. Points scored by the home team’s offense
9. Away team’s number of penalty yards
10. Home team’s yards per rush

***Bowls 2***

For the second collection of bowl games, the top team statistics shifted yet again, with the first three concerning tackling and rushing. Both away teams’ and home teams’ 3rd down conversion team statistics also saw some influence on game outcomes:

1. Home team’s opponents’ solo tackles for loss
2. Yards per rush allowed by the away team
3. Home team’s opponents’ tackle yards for loss
4. Away team’s scoring defense rank
5. Rush yards allowed by the away team
6. Away team’s fumbles lost rank
7. Away team’s yards per pass completion
8. Away team’s time of possession rank
9. Home team’s 3rd down conversion percentage rank
10. Away team’s 3rd down conversion percentage allowed

***Combined Dataset and Model Results***

As can be seen below, Table 2 contains information regarding the most influential team statistics over the course of the 2022 NCAA season:

**Table 2**

*Ten Most Influential Team Statistics: Combined*

|  |  |  |
| --- | --- | --- |
| Variable Name | Variable Explanation | Relative Importance (Mean Decrease in Accuracy) |
| Home\_Team\_Total\_Offense\_YPG | Home team’s offensive yards per game | 100.0 |
| Home\_Team\_Total\_Offense\_Yds/Play | Home team’s yards per play | 93.2 |
| Away\_Team\_Rushing\_Defense\_Yds/Rush | Yards per rush allowed by the away team | 82.1 |
| Away\_Team\_Rushing\_Defense\_Rank | Rushing defense rank of away team | 80.0 |
| Home\_Team\_Scoring\_Offense\_PPG | Home team’s points per game | 78.2 |
| Away\_Team\_Total\_Defense\_YPG | Away team’s yards allowed per game | 73.8 |
| Home\_Team\_Scoring\_Offense\_Rank | Home team’s scoring offense rank | 71.8 |
| Away\_Team\_Rushing\_Defense\_YPG | Rush yards allowed per game by away team | 70.5 |
| Home\_Team\_3rd\_Down\_Conversion\_Pct\_Pct | Home team’s 3rd down conversion percentage | 70.1 |
| Home\_Team\_Total\_Offense\_Rank | Home team’s total offense rank | 69.3 |

As with the other weeks, a random forest model was created in order to find the most influential variables in game outcomes. However, for the purposes of creating a college football predictive model, 20% of the data contained within the Combined Dataset was withheld in order to create a final test set. The random forest model was trained via 10-fold cross-validation on the 80% of data left in the training set. Accuracy statistics were collected following the assessment of the model’s performance on the test set. The test classification accuracy of the random forest model was 70.4%. This compares to the 56.8% accuracy that would be achieved by a naïve model. The sensitivity and specificity of the random forest model were estimated to be 76.1% and 63.0% respectively. Amongst the most important factors for the model were a home team’s offensive yards per game, a home team’s yards gained per play, and an away team’s yards allowed per rush.

**Discussion**

**Introduction**

As one might expect concerning a game as complex as NCAA football, the importance of various team statistics varied significantly over the course of the season. In the first weeks of the 2022 season, the top few statistics appeared to be defensively oriented. Yards per game allowed by the home team’s defense and rush yards allowed per game by the away team in weeks 2 and 3 were the most significant in predicting the outcomes of games in those weeks. This may suggest that early within the season, it is not the offense that is most important in determining game outcomes. Rather, the defense’s ability to limit the opposing offenses’ opportunities to gain yards, particularly through rushing plays, may be a worthy point of emphasis. Towards the midpoint of the 2022 football season, punt return yards, passing efficiency, as well as tackles for loss, and points per game were among the most influential factors in game outcomes. This may suggest that approaching the middle of the season, teams that excelled in a number of different areas of the game began to gain prominence, rather than teams that excel defensively toward the beginning of the season. At the end of the season (in championship games and bowl games), the most key team statistics that influenced game outcomes involved punt returns, kickoffs, and solo tackles for loss. Teams that excelled in these areas may have been more likely to qualify for and achieve success in those competitions. It may be the case that in general, special teams plays and defensive plays were what separated teams that qualified and won those competitions from teams that did not. Interestingly, though, the statistical results regarding the Combined Dataset for the 2022 season told a different tale. The two most influential variables that predicted outcomes throughout the season were offensively oriented. Offensive yards per game and offensive yards per play were most influential here. While other team statistics were found to be more significant in college football game outcomes at various points throughout the season, it may be the case that, overall, emphasizing offensive firepower in a team is the surest path to success throughout the course of an entire season.

**Recommendations**

Of course, it is important to take this work within its proper context. As the analysis that was conducted was focused on providing a case study of the 2022 NCAA season, the findings of this work will be somewhat limited outside of that season. Each analysis of team statistics within the weeks included in this work serves as a sort of snapshot in time of the most influential team statistics. Therefore, it would be ill-advised to apply the results of this analysis except for, perhaps, the statistical results of the Combined Dataset to other seasons of college football. Given that that dataset contained significantly more observations than the rest of the datasets analyzed, it likely provides more statistical power in terms of making predictions outside the 2022 season than the others. Nonetheless, the game of college football is not only subject to significant variance in terms of team statistics that are influential in any given week, but also subject to rule changes and paradigm shifts over the course of time. Also, unlike various other academic works in this area, considerations such as individual player statistics and intra-play statistics were not included in the overall analysis. Thus, it is recommended that the results of this work be combined with the aforementioned statistical areas in future works in order to provide a fuller picture of not only the 2022 season, but the game of NCAA football as a whole.

**Conclusion**

The purpose of this work was to serve as a case study of the 2022 college football season. Pursuant to that purpose, a broad overview of the history of college football was provided, relevant academic works in this area were summarized, and an analysis of several weeks of the season was conducted. Along with this, a random forest model was created in order to determine the most prominent team statistics that played a role in game outcomes. The results of this analysis indicated a wide variance in influential team statistics on a weekly basis through the 2022 season. However, more stable team statistics, including total offense yards per game and offensive yards per play were found to be most influential over the course of the entire season. A random forest model analysis found that these and eight other team statistics found in Table 2 were the most significant in predicting 70.4% of games analyzed. It is recommended that caution be applied to generalizing the weekly results of this analysis and that other statistical areas of the game including intra-play and individual player statistics be complimented with the results of this analysis in order to provide a fuller grasp on the 2022 NCAA football season and beyond.

**References**

Beaton, A. (2016, January 11). The Count: How Much Is Your Favorite College Football Team

Worth? *Wall Street Journal*. Retrieved April 28, 2023, from

https://www.proquest.com/docview/1755383059/5C5EF1F010814C19PQ/1?accountid=9

358.

*College football playoff payouts 2022-2023*. Business of College Sports. (2022, November 23).

Retrieved April 28, 2023, from https://businessofcollegesports.com/college-football-

playoff-payouts/

Fair, R. C., & Oster, J. F. (2002, October 10). *College football rankings and market efficiency*.

SSRN. Retrieved March 7, 2023, from

https://papers.ssrn.com/sol3/papers.cfm?abstract\_id=335801

Fox. (2006). *BCS Chronology*. Fox Sports on MSN - BCSFOOTBALL - BCS Chronology.

Retrieved April 10, 2023, from

https://web.archive.org/web/20070915010233/http://www.bcsfootball.org/bcsfb/history

Gay, J. (2018, December 17). College Football: Expand the Playoff, College Football --- The

current format has sucked all the urgency out of bowl season. But money is what will

push the playoff to eight teams. *Wall Street Journal*. Retrieved April 28, 2023, from

https://www.proquest.com/docview/2157289587/citation/F6CFA6E965434580PQ/1?acc

ountid=9358.

Huddleston, T. (2020, September 2). *College football stars could be earning as much as $2.4*

*million per year, based on NCAA revenues: Study*. CNBC. Retrieved April 28, 2023,

from https://www.cnbc.com/2020/09/02/how-much-college-athletes-could-be-earning-

study.html?&qsearchterm=NCAA+revenue

Kleinman, D. (2019, October 15). *The most valuable college apparel deals: UCLA leads as Gear*

*Companies' new mindset thwarts rivals*. Forbes. Retrieved April 28, 2023, from

https://www.forbes.com/sites/danielkleinman/2019/09/13/ucla-most-valuable-college-

apparel-deals/?sh=7978555f6762

Leung, C. K., & Joseph, K. W. (2014). Sports data mining: Predicting results for the college

football games. *Procedia Computer Science*, *35*, 710–719.

https://doi.org/10.1016/j.procs.2014.08.153

LoPilato, A. C., Hoffman, B. J., & Overstreet, B. L. (2013). Outcomes of peak, typical, and

variability in performance of college football teams. *Journal of Business and Psychology*,

*29*(2), 221–233. https://doi.org/10.1007/s10869-013-9336-3

Mankin, J. A., Rivas, J., & Jewell, J. J. (2021). The Effectiveness of College Football Recruiting

Ratings in Predicting Team Success: A Longitudinal Study. *A Longitudinal Study.*

*Research in Business and Economics Journal*, *14*, 4–22.

Mullaney, T. (2021, June 23). *Supreme Court NCAA ruling and the New Future of paying*

*college athletes*. CNBC. Retrieved April 28, 2023, from

https://www.cnbc.com/2021/06/21/supreme-court-ncaa-decision-how-college-athletes-

plan-to-cash-in.html

Pankoff, L. D. (1968). Market efficiency and football betting. *The Journal of Business*, *41*(2),

203. https://doi.org/10.1086/295077

The Professional Football Researchers Association. (n.d.). Camp and his Followers American

Football 1876-1889. Retrieved April 10, 2013, from

https://web.archive.org/web/20100929152206/http://profootballresearchers.org/Articles/

Camp\_And\_Followers.pdf

Saul, D. (2022, January 12). *Viewership for college football playoff championship up from*

*record low 2021 - but still below NFL's ratings*. Forbes. Retrieved April 28, 2023, from

https://www.forbes.com/sites/dereksaul/2022/01/11/viewership-for-college-football-

playoff-championship-up-from-record-low-2021---but-still-below-

nflsratings/?sh=62cf760e3622

Schafer, J. (2022, August 12). *Fanduel posts profitable quarter despite leaning in 'harder' on*

*customer acquisition*. Yahoo! Finance. Retrieved April 28, 2023, from

https://finance.yahoo.com/news/fanduel-profitable-quarter-customer-acquisition-

earnings170811528.html?guccounter=1&guce\_referrer=aHR0cHM6Ly93d3cuYmluZy5j

b20v&guce\_referrer\_sig=AQAAAJeueE7Dbpg\_\_2qWawyjqOiyAtrjloMlW7MVOC4cB

hUoxg9Z0QnlJ6301SgmLc8tDB-hUY3

oTNuKEkJHXMt6fu4zSeATOyS6T9sjBRMpcXTMUe4VFoG284eXXYN1SmLUwPuD

HGnZXUzTDJ--1r2U96xmp04kZqUAOl4IzTi8e7O

Sports betting in America is exploding: Daily Chart. (2022, February 10). *The Economist*.

Retrieved April 28, 2023, from

https://www.proquest.com/docview/2627821387/citation/CB05226F39644FC9PQ/1?acc

ountid=9358.

Sports Business Journal. (2023, April 25). *College football playoff spring meeting kicks off*.

Sports Business Journal. Retrieved April 28, 2023, from

https://www.sportsbusinessjournal.com/Daily/Morning-Buzz/2023/04/25/college-

football-playoff-spring-meeting.aspx?hl=college%2Bfootball&sc=0

Stefani, R. (1980). Improved least squares football, basketball, and soccer predictions. *IEEE*

*Transactions on Systems, Man, and Cybernetics*, *10*(2), 116–123.

https://doi.org/10.1109/tsmc.1980.4308442

Stern, H. (1991). On the probability of winning a football game. *The American Statistician*,

*45*(3), 179. https://doi.org/10.2307/2684286

Vergin, R. C., & Scriabin, M. (1978). Winning strategies for wagering on National

Football League games. *Management Science*, *24*(8), 809–818.

https://doi.org/10.1287/mnsc.24.8.809

West, B. T., & Lamsal, M. (2008). A new application of linear modeling in the prediction of

college football bowl outcomes and the development of team ratings. *Journal of*

*Quantitative Analysis in Sports*, *4*(3). https://doi.org/10.2202/1559-0410.1115

Wilson RL (2020) College Football Overtime Outcomes: Implications for In-Game Decision-

Making. *Front. Artif. Intell.* 3:61. doi: 10.3389/frai.2020.00061

Yurko, R., Matano, F., Richardson, L. F., Granered, N., Pospisil, T., Pelechrinis, K., & Ventura,

S. L. (2020). Going deep: Models for continuous-time within-play valuation of game

outcomes in American football with tracking data. *Journal of Quantitative Analysis in*

*Sports*, *16*(2), 163–182. https://doi.org/10.1515/jqas-2019-0056

Zuber, R. A., Gandar, J. M., & Bowers, B. D. (1985). Beating the spread: Testing the

efficiency of the gambling market for National Football League games. *Journal of*

*Political Economy*, *93*(4), 800–806. https://doi.org/10.1086/261332

**Appendix**

Code and Materials

Code and materials utilized in this work can be found using the following link: https://www.dropbox.com/sh/3mf7e11dkyl4ddn/AABPuwaETD\_oGfuxAsrYZdsra?dl=0