MODELING WINNING TEAMS FOR 2019 NCAA DIVISION I FBS FOOTBALL SEASON

by

DE'QUANTE MCKOY

A thesis submitted to the Graduate Faculty of
Elizabeth City State University
in partial fulfillment of the
requirements for the Degree of
Master of Science in Mathematics

May

2020

APPROVED BY

Julian A. D. Allagan, Ph.D. Committee Chair	Kenneth L. Jones, Ph.D. Committee Member
Dipendra Sengupta, Ph.D. Committee Member	Gabriela H. Del Villar, Ph.D. Committee Member

ABSTRACT

Given the 2019 NCAA football data, this thesis explores the effect of several variables such as opponent 3rd down Conversion, Turnovers, and Point per Play Margin, etc., on the Win percent or Win margin percent for each team. We run both logistic and linear regression models and found Point Per Play Margin to be consistently statistically significant at 5% risk level. With a linear model on a continuous Win margin, Point Per Play Margin and Opponent 3rd Down Conversion were statistically significant at explaining 90% of the variations in the response. However, with a logistic regression model on a binary win margin, Point Per Play Margin was the only statistically significant variable at classifying the response Win margin (above 50% or not) with an accuracy rate of 86%.

DEDICATION

I want to dedicate this thesis to my family for being able to help me mentally through these two years of graduate school. I would like to dedicate this research to my little brother who gave me the idea to see if I would be able to predict the next winner of the NCAA College football championship winner. Finally, I would like to give a special dedication to my mother without her support and mental, emotional guidance I would not be here today.

ACKNOWLEDGEMENT

My sincere and utmost gratitude goes to the department of Mathematics. I would like to thank my peers and also the department chair Dr. Kenneth L Jones. I would like to give thanks to my my advisor Dr.Allagan through his perspective guidance throughout my thesis research. I would like to thank all the professors over the span of these past two years of graduate school that help mold me to this point of preparing to defend my thesis to the committee board. Finally, I would like to give thanks to my family who prayed for me and supported me mentally and physically throughout my time in the graduate program at Elizabeth City State University.

Contents

List of Figures

Chapter 1 Introduction

1.1 Overview

For this thesis, we use excel to clean or adjust the content of the data and then, we use SPSS (Statistical Package for Social Sciences) software for our analysis. In this chapter (Chapter 1), we present a detailed background of the history of National Collegiate Athletic Association (NCAA) football. In Chapter 2, we present several rules and plays of the football game. In Chapter 3 we introduce the readers to some fundamental statistical notions, describe the data and present several statistics on the variables in the data. In Chapter 4, we perform several analysis. We run both logistic and linear regression models. With a linear model on a continuous win margin, Point Per Play Margin and Opponent 3rd Down Conversion were statistically significant at explaining 90% of the variations in the response. However, with a logistic regression model on a binary win margin, Point Per Play Margin was the only statistically significant variable at classifying the response win margin (above 50% or not) with an accuracy rate of 86%. We conclude this thesis (Chapter 5) with some of the issues we run into with the data and make some recommendation for a future research direction.

1.2 History of NCAA football

The birth of American football came in 1869 on College Avenue in New Brunswick, New Jersey. The game was between Rutgers University and the College of New Jersey (now known as Princeton University). There were 25 players on the field for both teams and the rules were based on the London Football Association, which did not

allow players to either pick up or throw the ball. The game resembled a form of soccer or rugby — something that if viewed in the context of football today, would look like one extended fumble with players trying to kick or hit the ball across the opposing team's goal line. The game resulted in a 6-4 victory for Rutgers and attracted around 100 spectators. Even after the emergence of the professional National Football League (NFL), college football remained extremely popular throughout the U.S.[2] Although the college game has a much larger margin for talent than its pro counterpart, the sheer number of fans following major colleges provides a financial equalizer for the game, with Division I programs — the highest level — playing in huge stadiums, six of which have seating capacity exceeding 100,000 people. In many cases, college stadiums employ bench-style seating, as opposed to individual seats with backs and arm rests (although many stadiums do have a small number of chair-back seats in addition to the bench seating). This allows them to seat more fans in a given amount of space than the typical professional stadium, which tends to have more features and comforts for fans. (Only three stadiums owned by U.S. colleges or universities — Cardinal Stadium at the University of Louisville, Georgia State Stadium at Georgia State University and FAU Stadium at Florida Atlantic University — consist entirely of chair-back seating).College athletes, unlike players in the NFL, are not permitted by the NCAA to be paid salaries. Colleges are only allowed to provide non-monetary compensation such as athletic scholarships that provide for tuition, housing, and books. Early games appear to have had much in common with the traditional "mob football" played in Great Britain. The games remained largely unorganized until the 19th century, when intramural games of football began to be played on college campuses. Each school played its own variety of football. Princeton University students played a game called "ballown" as early as 1820. A Harvard tradition known as "Bloody Monday" began in 1827, which consisted of a mass ballgame between the freshman and sophomore classes. In 1860, both the town police and the college authorities agreed the Bloody Monday had to go. The Harvard students responded by going into mourning for a mock figure called "Football Fight", for whom they conducted funeral rites. The authorities held firm and it was a dozen

years before football was once again played at Harvard. Dartmouth played its own version called "Old division football", the rules of which were first published in 1871, though the game dates to at least the 1830s. All of these games, and others, shared certain commonalities. They remained largely "mob" style games, with huge numbers of players attempting to advance the ball into a goal area, often by any means necessary. Rules were simple, violence and injury were common.[4][5] The violence of these mob-style games led to widespread protests and a decision to abandon them. Yale, under pressure from the city of New Haven, banned the play of all forms of football in 1860. American football historian Parke H. Davis described the period between 1869 and 1875 as the 'Pioneer Period'; the years 1876–93 he called the 'Period of the American Intercollegiate Football Association, and the years 1894–1933 he dubbed the 'Period of Rules Committees and Conferences. Columbia University was the third school to field a team. The Lions traveled from New York City to New Brunswick on November 12, 1870 and were defeated by Rutgers 6 to 3. The game suffered from disorganization and the players kicked and battled each other as much as the ball. Later in 1870, Princeton and Rutgers played again with Princeton defeating Rutgers 6-0. This game's violence caused such an outcry that no games at all were played in 1871. Football came back in 1872, when Columbia played Yale for the first time. The Yale team was coached and captained by David Schley Schaff, who had learned to play football while attending Rugby School. Schaff himself was injured and unable to the play the game, but Yale won the game 3-0, nonetheless. Later in 1872, Stevens Tech became the fifth school to field a team. Stevens lost to Columbia but beat both New York University and City College of New York during the following year. By 1873, the college students playing football had made significant efforts to standardize their fledgling game. Teams had been scaled down from 25 players to 20. The only way to score was still to bat or kick the ball through the opposing team's goal, and the game was played in two 45-minute halves on fields 140 yards long and 70 yards wide. On October 20, 1873, representatives from Yale, Columbia, Princeton, and Rutgers met at the Fifth Avenue Hotel in New York City to codify the first set of intercollegiate football rules. Before this meeting, each school had its own set of rules and games

were usually played using the home team's own code. At this meeting, a list of rules, based more on the Football Association's rules than the rules of the recently founded Rugby Football Union, was drawn up for intercollegiate football games.

The game we in the United States know as football is all the more appropriately called turf football, for the vertical yard lines that mark the field. Firmly identified with two English games—rugby and soccer (or affiliation football)— turf football started at colleges in North America, essentially the United States, in the late nineteenth century. On November 6, 1869, players from Princeton and Rutgers held the main intercollegiate football challenge in New Brunswick, New Jersey, playing a soccer-style game with rules adjusted from the London Football Association. While various other first class Northeastern schools took up the game during the 1870s, Harvard University kept up its separation by adhering to a rugby-soccer half breed called the "Boston Game." In May 1874, after a match against McGill University of Montreal, the Harvard players chose they favored McGill's rugby-style rules to their own. In 1875, Harvard and Yale played their first intercollegiate match, and Yale players and observers (counting Princeton understudies) held onto the rugby style too. [6]

The man generally answerable for the change from this rugby-like game to the game of football we realize today was Walter Camp, known as the "Father of American Football." As a Yale undergrad and clinical understudy from 1876 to 1881, he played halfback and filled in as group chief, proportionate to lead trainer at that point. Considerably more significantly, he was the controlling power on the standards leading group of the recently framed Intercollegiate Football Association (IFA). Because of Camp, the IFA made two key advancements to the youngster game: It got rid of the opening "scrummage" or "scrum" and presented the necessity that a group surrender the ball subsequent to neglecting to descend the field a predetermined yardage in a specific number of "downs." Among different developments Camp presented were the 11-man group, the quarterback position, the line of scrimmage, hostile sign calling and the scoring scale utilized in football today. Notwithstanding his work with the guidelines board, Camp instructed the Yale group to a 67-2 record from 1888 to

1892—all while filling in as an official at a watch-producing firm. [6]

Chapter 2 Basic College Football Game Terminologies

Here, we introduce the reader to the rules and regulations, and, some common plays of the football game.

2.1 Rules

1. **Backfield:** This is the area of the football field behind the Offensive Line where a maximum of four (4) offensive players can stand on any given play. It is in the opposite direction that the ball is moving in.

2. **Downfield:** Toward or in the defensive team's end of the playing field. From the offense's point of view, it is past the Line of Scrimmage and at or toward the goal line of the defensive team. [8]

3. **Fair catch:** This happens when the player who is catching the kickoff or punt signals to the referee that he is not going to run with the ball once he catches it, but rather accepts field position where he is. The defensive players are not allowed to tackle him at this point, but they still cover the play in the event he drops the ball because then the ball becomes "live" again.

4. **Field goal:** This is worth three points. During a field goal attempt the ball is snapped from where it was last spotted, and the kicker stands straight back from it. In order to determine how long a field goal attempt is you need to consider the distance from the beginning of the end zone to the goal posts (10 yards) as well as how far the kicker stands from where the ball is snapped (7 yards). This is an additional 17 yards in total. Therefore, if the ball is on an opponent's 30-yard line, then the field goal will be a 47-yard attempt. Especially

in college, an offense wants to position the ball so that the field goal kicker has less than about 40 yards for the attempt, which means they really aim to be within the opponent's 23 yards line. Field goals are attempted on 4th down because they have a higher likelihood of putting points on the board than going for a touchdown. [8]

5. **Forward pass:** This is when the football is thrown by a football player (typically the quarterback) to one of the players on his team (typically the wide receiver or tight end but can also be to the running back and if specifically declared eligible before the play is run, can also include the offensive linemen and the quarterback). Only one forward pass can be thrown by the football team during a given play and it must be thrown from behind the line of scrimmage (where the ball begins).[8]

6. **Goal post:** The structure in the end zone that the kicker must kick the football through to score an extra point or a field goal. Hash Marks (Inbound Lines) In college these are the marks on either side of that are 20 yards from the sidelines (please note that this differs significantly from the NFL). The width of the field is a total of 53 yards and therefore the left and right hash marks are separated by 13 yards. The play will always begin on or within the hash marks, which run the width of the field. If the ball ends up within the hash marks after a play, the ball will be spotted at that exact position. However, if the ball ends up outside the hash marks, it will be moved directly left or right so that it is placed on the closest hash mark after the play. [6]

7. **Lateral:** A pass that is thrown backward by the team with the ball (meaning in the opposite direction that they are attempting to move). While players can attempt only one forward pass during a given play, they can lateral a pass as many times as they would like. A play with multiple lateral passes is most often seen on the last play of the game when a team has a lot of ground to cover and needs to score a touchdown to tie or win the game.

8. **Line of scrimmage:** Each team has its own line of scrimmage when the ball is

ready for play, which is the yard line and its vertical plane that passes through the point of the ball nearest its side of the field and extends to the sidelines. [6]

9. **Neutral zone:** The space between the two lines of scrimmage (on offense and defense) extended to the sidelines. It is the length of the ball. Only the player snapping the ball can be in this area once he puts his hands on the ball or simulates such. If an offensive player enters the neutral zone after this point, it is encroachment. If a defensive player enters the neutral zone after this point, it is offside. There are penalties for both infractions.

10. **Penalty:** This is prescribed for a rule infraction. If there is a penalty for the rule infraction it is called a foul. If there is no penalty for the rule infraction, it is simply a violation. A violation cannot offset a foul. However, fouls can offset each other. [6]

11. **Personal Foul:** This type of foul typically carries a 15-yard penalty plus an automatic first down if committed by the defense. It includes such penalties such as Blocking Below the Waist and Chop Block. [7]

12. **Punt:** This takes place on 4th down, which is the last down of possession before the ball goes to the other team. Rather than call a play where they run or pass the ball, the team on offense has the opportunity to kick the ball to the other side. If a team is close to their own end zone, they will likely opt for this play so that they move the ball in the opposite direction. The risk is when the other team has a good punt returner because he can negate the benefit of kicking the ball and at worst can score a touchdown on the punt return. But if executed correctly, this puts the punting team in a much better position than if they were to lose the ball after 4 downs and just give the other team the ball at the spot on the field where they had it. You can distinguish this type of kick from a kickoff because the ball is snapped to the kicker who holds the ball in his hands and drops it to kick, rather than kicking it from the ground, where it sits on a tee. [6]

13. **Red-shirt:** The designation given to a college player who did not play in any games during a particular year due to the Head Coach's decision or injury (sometimes in the case of injury the NCAA will grant a medical red-shirt to a player who started the season but did not play in very many games); The red-shirted player is permitted to practice with the team and this doesn't count against his four years of eligibility. Most often freshman is red-shirted so that in their first year of playing, which is their sophomore year of college, they are called "red-shirt freshmen" as opposed to "true freshman". [6]

14. **Red Zone:** This is a term used to describe the area between one's opponent's 20-yard line and its goal line. The chance of scoring is increased greatly once a team reaches the Red Zone because even if they don't make it into the End Zone, at that point the field goal attempt would be 37 yards (20 +17) or less and it is expected that the kicker will be able to make it successfully. [7]

15. **Sack:** This occurs when the defense is able to tackle the quarterback behind the line of scrimmage during a passing play. If the play is designed as a running play it is not a Sack, but rather a Stuff. A Sack only occurs when the Quarterback is attempting a passing play and is tackled. In College Football the negative yards for a Sack are subtracted from the Quarterback's rushing totals and therefore the team's total rushing yards, even though the Quarterback was attempting to throw the ball. This is why the total passing and receiving yards in College Football are identical. However, these numbers are different in the NFL because the negative yards from a Sack are subtracted from the Quarterback's and therefore the team's total passing yards and thus that total will differ from the number of receiving yards.[8]

16. **Safety:** (1) When the player holding the football is tackled in his own end zone. The defense gets 2 points plus possession of the football via a kickoff. [7]

    (2) A type of Defensive Back, which can be either a Free Safety or Strong Safety. See Defensive Back.

17. **Series:** A series comprises four consecutive downs that each begins with the

snap.

18. **Snap:** The center hands off the football between his legs to a player standing behind him (usually the quarterback) at the start of each play. This can also be referred to as a hike or hiking the ball as opposed to snapping it.[6]

19. Special teams: This is the unit on the team involved in any of the kicking plays, including the kickoff, the punt and a field goal attempt.

20. **Stuff**: A tackle of a ball carrier on a running play, behind the line of scrimmage.

21. **Touchdown:** This is when a player in possession of the football takes the ball across the plane of the other team's goal line and into the end zone while remaining in bounds. It is worth 6 points plus the option to run one additional play, which is either to try to kick the ball through the goal posts for one extra point, or to try to get the ball into the end zone again for two extra points. [6]

22. **Towels:** Many people don't realize this, but there are actually rules when it comes to the towels some players are wearing on the field. On scrimmage plays (meaning non-kick-off plays), one white moisture-absorbing towel may be worn by one interior offensive lineman, one offensive backfield player and a maximum of two defensive players. The towels of the offensive backfield and defensive players must be 4 inches by 12 inches and must be worn on the front or side of the belt. There are no restrictions on the size or location of the towel worn by the interior offensive lineman; On free kicks, one white moisture absorbing towel without markings may be worn by a maximum of two kicking team and two receiving team players. The towels worn on free kicks must be 4 inches by 12 inches and must be worn on the front or side of the belt.

23. **Turnover:** When the offense, the team with the possession of the football, loses it and "turns" it over to the other team either through a fumble (dropping the ball or otherwise losing it to a defensive player) or an interception (when a defensive player catches the ball in the air).[7]

24. **Up-field:** Toward or in the offensive team's end of the playing field. From the defense's point of view, it is past the Line of Scrimmage and at or toward the goal line of the offensive team. [6]

2.2 Penalties

Definition: When a foul has been committed (a rule infraction that carries a penalty) a penalty will be enforced, offset or declined. These are some of the common penalties and their respective Referee signals.

1. **Block in the Back:** An illegal Block in the Back involves contact against an opponent (other than the ball carrier) occurring when the force of the initial contact is from behind and above the waist. (Contrast this with Clipping which is from behind and at or below the waist). An easy way to remember this is that the Block in the Back happens in a player's back. There are some exceptions, including when a player is attempting to tackle a runner or when a player is attempting to recover a ball. The penalty is 10 yards.[6]

2. **Blocking Below the Waist:** Blocking below the waist is the initial contact below the waist with any part of the blocker's body against an opponent (other than the ball carrier). This applies to blocking an opponent who has one or both feet on the ground. A blocker who makes contact above the waist and then slides below the waist is considered to have blocked Above the Waist and not Below the Waist. Blocking Below the Waist and from behind is never permitted. Otherwise, Blocking Below the Waist from the front may be permitted depending on the rules, which are rather lengthy and can be found at Rule 9, Section 1, Article 2(e) of the Football 2009-2010 Rules and Interpretations. If it happens to be one of the blocks Below the Waist that is not permitted, this is a personal foul that includes a 15-yard penalty and an automatic first down if committed by the defense.

3. **Chop Block:** A combination block by any two players against an opponent (other than the ball carrier) with or without delays between the blocks in which one player blocks low and the other blocks high. It can come in as a high then low block or a low than high block. This is a personal foul and the penalty is 15 yards plus an automatic first down if the foul was committed by the defense.

4. **Clipping:** A block against an opponent (other than the ball carrier) occurring when the force of the initial contact is from behind and at or below the waist. (Contrast this with a Block in the Back that is from behind but above the waist). There are some exceptions, including when a player is attempting to tackle a runner or when a player is attempting to recover a ball. This is a personal foul and the penalty is 15 yards plus an automatic first down if the foul was committed by the defense.

5. **Encroachment:** A penalty when an offensive player other than the Center (or whoever is snapping the ball) is in or beyond the neutral zone once the snapper touches or simulates touching the ball before the snap.

6. **False Start:** When an offensive player pretends to charge forward or shifts/moves in a way that simulates the beginning of a play. Once the snapper assumes the position for snapping the ball and touches/simulates touching it, if he moves to another position, this is considered a false start. It also includes quick movement by any of the four other Offensive Linemen other than the snapper (wearing number 50 though 79) after having placed a hand/s on or near the ground (unless the movement was a reaction to the defensive player's movement into the neutral zone). The penalty for fouls before the ball is snapped is five yards from the succeeding spot.

7. **Holding:** This is one of the most popular penalties in football. It occurs when a player on offense or defense stops the movement of his opponent by holding onto his body (which includes his uniform). The penalty is 10 yards. (In the NFL it is a loss of 10 yards unless committed by the defense and then it is 5 yards and an automatic first down).

8. **Illegal Touching:** When a player who is ineligible intentionally touches a ball that has been kicked or passed before an opponent or official touch it. If this occurs on a forward pass and the player is ineligible to touch it, there is a five-yard penalty from the previous spot. [7]

9. **Intentional Grounding:** A penalty when the quarterback purposely throws an incomplete pass just to avoid a sack. In college the penalty is that the ball is then marked at the spot of the foul and there's a loss of down, which is exactly what would have happened had the play run without the penalty and the Quarterback had been sacked. So there really is no consequence for the Quarterback to attempt to get away with avoiding the Sack. However, in the NFL the rules are different. It's a 10-yard penalty plus loss of down, unless the Quarterback is farther behind the line of scrimmage than 10 yards, then it's at the spot of the foul. Regardless, if this play occurs in a player's own End Zone, then the result of the play is a Safety.

10. **Loss of a Down:** This is an abbreviation meaning "loss of the right to repeat a down." Oftentimes after penalties, a team on offense will have the right to repeat the down, but on certain penalties they lose this right. A well-known broadcaster once had this very wrong during the broadcast and mistakenly thought after a penalty for intentional grounding that took place on 1st down that the loss of down penalty meant that rather than 2nd down, it was 3rd down. He proceeded to do an analysis on this that was completely wrong as the fact is that a team never actually loses a down, just the right to repeat it.

11. **Offside:** After the ball is ready for play, when a defensive player (a) contacts an opponent beyond the Neutral Zone before the ball is snapped, (b) contacts the ball before it is snapped, (c) threatens an Offensive Linemen such that it causes an immediate reaction before the ball is snapped or (d) is in or beyond the Neutral Zone when the ball is legally snapped. The penalty is 5 yards. [6]

12. **Pass Interference:** This can be a foul committed by either the offense or defense. The rules on this are quite complicated as written and more in-depth

than I would have thought, but the general idea is that when there are two players in the vicinity of the ball, they need to be going after the ball and not the other player. The penalty for this is 15 yards plus an automatic first down if committed by the defense.

13. **Personal Foul:** This type of foul typically carries a 15-yard penalty plus an automatic first down if committed by the defense. It includes such penalties such as Blocking Below the Waist and Chop Block.

2.3 Plays

2.3.1 General Plays

1. **Passing Plays:** When a passing play occurs, the backs and receivers run specific patterns, or routes, and the quarterback throws the ball to one of the players. On these plays, the offensive line's main job is to prevent defensive players from tackling the quarterback before he throws the ball (a "sack") or disrupting the quarterback in any other way during the play.

2. **Run Plays:** A running play occurs when the quarterback hands the ball to another player, who then attempts to carry the ball past the line of scrimmage and gain yards, or the quarterback keeps the ball himself and runs beyond the line of scrimmage. In both cases, the offensive line's main job is to run block, preventing the defensive players from tackling the ball carrier.

3. **Dive Play:** The Dive Play is the most basic running play in football. The dive requires a running back to take the hand-off and "dive" into the center of the line. A dive can take place between a center and guard or between a guard and a tackle.

4. **Off Tackle Run play:** Off Tackle plays means the running back runs off the outside hip of the tackle. The off-tackle run is probably the most-used play in football.An Off Tackle play gives the running back more room to maneuver than

a dive play, and the runner has the chance to pick his hole along the line, either cutting upfield to get into the defensive backfield quickest and get quick yards, or cutting outside where there should be more room to run and a potentially bigger play.

5. **Trap Play:** The trap play is where a guard from one side of the line runs parallel to the line of scrimmage and blocks on the other end of the line. The pulling guard is likely to blindside the end or outside linebacker he is blocking, creating a hole in the defense. The danger of trap plays is it tends to leave two defenders momentarily uncovered: the defensive lineman whom the guard usually blocks and the defensive end/linebacker the guard is trapping. The fullback often "seals" the play by blocking the defensive linemen the guard leaves uncovered.[7]

2.3.2 Offensive Terminology

1. **Cut or Cut-back:** This is when the player with the ball makes a sudden change in direction that makes it more difficult for defenders to tackle him.

2. **Double Option:** This is an offensive play in which there are two different options for who can run with the football and the Quarterback must make the decision on what to do with the ball based on the Zone Read.

3. **Draw Play:** A play in which the quarterback pretends that he is going to pass the ball and drops back as if to pass in order to draw the defenders downfield into pass defense. The Quarterback then either hands the ball off to a Running back or keeps it and runs with it himself.

4. **Eligible Receiver:** Only certain players on offense are permitted to catch a forward pass. No player wearing number 50 through 79 is permitted to catch a forward pass (these are the numbers designated and required to be worn by at least 5 Offensive Linemen). As long as the players on the end of the Offensive Line are not wearing numbers 50-79, they are eligible to catch a forward pass.

All defensive players are permitted to touch or catch a pass. Once a defensive player touches a legal forward pass, all players become eligible.

5. **End Around:** A play where the Wide Receiver moves into the backfield as the ball is snapped and takes the handoff directly from the quarterback. He then runs around the opposite end from where he originally lined up prior to the snap. This is different from a Reverse in which the Quarterback hands the ball to another player who then hands it to the Wide Receiver.

6. **Flanker:** A wide receiver who lines up in the backfield outside of another receiver. He is also called the "Z" Receiver.

7. **Flea Flicker:** A trick play on offense in which the Quarterback hands the ball to the Running Back who throws a backward pass back to the Quarterback, who then throws a forward pass to a Wide Receiver or Tight End.

8. **Hurry-Up Offense:** An offensive strategy designed to quickly carry out offensive plays while using as little time off of the clock as possible. It often involves going without a huddle prior to the plays and can be difficult for a defense as they have less time to get into position and recognize what the offense is doing.

9. **Option:** This is an offensive scheme that has as its premise two or more options of what the Quarterback can do with the ball based on what he sees from the defense. His decision is based on the Zone Read. See Spread Option, Triple Spread Option and Zone Read. Pass Protection: Blocking by the offensive football players to keep defenders away from the Quarterback to give him time to throw the football.

10. **Pistol Formation:** A hybrid version of the shotgun in which the Quarterback lines up about 3 yards behind the Center and the Running Back lines up directly behind the Quarterback. The Pistol puts the quarterback in the Shotgun, but only 3 or 4 yards behind the Center instead of 5 to 7 yards back. But unlike a typical Shotgun in which the Running Back is next to the Quarterback, the Pistol puts the Running Back behind the Quarterback. This gives

the Quarterback the time and vision needed for the passing game while letting the Running Back get moving toward the Line of Scrimmage so that if he takes the handoff it is while moving as opposed to standing still. "Your back now has the ability to go both ways, as opposed to being offset one way or the other," (Ohio State Head Coach Jim Tressel). The formation also provides an element of deception, with the Running Back almost hiding behind the Quarterback, so opposing Linebackers can't get a read on the run play. The Option can be run out of this formation as well.

11. **Play Action Pass:** This is when the Quarterback pretends to hand off the football to the Running Back, but actually keeps it. The Running Back will run up the field and pretend that he has the ball. The Offensive Linemen will join in the fake and act as if they're blocking for a running play. But it's all just an attempt to disguise the pass play, as the Quarterback never actually gave the ball away and is actually attempting to pass it. The offense is hoping that the defensive players will react to what they think is a running play by moving up to defend the runner who is pretending to have the ball, rather than continue with the pass rush or with covering the wide receivers. If executed correctly this will give the quarterback and receivers more time and space to make a play. This is the opposite of a Draw Play.

12. **Pocket:** The area surrounding the Quarterback where he stands when the ball is snapped. It's really just the space around the quarterback where that he hopes is protected from the defense by his blockers. When this "collapses" it means that the defensive players have broken through and the Quarterback no longer has the protection of his Pocket to stand in and pass the ball so he must move immediately.

13. **Reverse:** An offensive play in which the Quarterback hands the ball to the Running Back who starts by carrying the ball toward one side of the field, but then hands or tosses the ball to a teammate (almost exclusively a Wide Receiver) who is running in the opposite direction. This is in contrast to an End Around

in which the Quarterback hands the ball directly to the Wide Receiver. Run out of the Gun: When a running play takes place from the Shotgun Formation.

14. **Screen Pass:** A short pass thrown to a receiver on the outer edge of the field while he stands behind one of his Offensive Linemen, who is standing in between him and the defensive player. Screens are used constantly in basketball and while they are complicated to describe in writing, they are much easier to watch. So just watch and listen.

15. **Formation:** This is an offensive formation in which the Quarterback lines up about 5 to 8 yards back from the Center when receiving the snap as opposed to directly behind him. He is often (but not necessarily) accompanied by one or two Running Backs standing directly next to him. One of the reasons why an offense will use a Shotgun Formation is so the Quarterback can more easily see the field and scan from left to right and see where the defensive players are coming from. The Quarterback needs this vision when executing passing plays and that is why this formation is typically associated with such. However, when it is not a passing play, it is considered a "Run out of the Gun." [6]

16. **Splits:** The distance between the feet of adjacent Offensive Linemen. It is said to be wide, if there is a large gap between players, or narrow, if the gap is small.

17. **Spread Offense:** The "spread" refers to any formation that forces the defense to cover more area before the play begins (before the ball is snapped) and creates more area in between defensive players. Because the Offensive Tackles (one on each end of the offensive line) stand farther away from each other, the Defensive Linemen and Linebackers must mimic their counterparts and spread out as well.

18. **Spread Option:** This is an offensive scheme that incorporates the Spread Offense and the Option. It has many different variations and is notably implemented at Florida by Urban Meyer, at Michigan by Rich Rodriguez and at Oregon by Chip Kelly (and formerly by Mike Bellotti).

19. Triple Option: This is an offensive play in which there are three different options for who can run with the football and the Quarterback must make the decision on what to do with the ball based on the Zone Read. See Option, Triple Spread Option.

20. **Triple Spread Option:** Also referred to as the Triple Option Spread Offense, this is a version of the Spread Option Offense that is credited to Paul Johnson who installed it at Navy and currently runs it at Georgia Tech. It is distinguished form the other Spread Options in that it uses three different running options on each play. The Quarterback makes the decision on what to do with the ball based on the Zone Read. Wildcat Formation: An offensive formation for a running play in which the runner takes the snap directly from the Center.

21. **Zone Read:** This is a step in the Spread Option in which the Quarterback reads the defense in order to determine which option to use. The terms Spread Option and Zone Read are used interchangeably in the broadcast as both sufficiently describe the offensive strategy because the Zone Read and Option work in conjunction with each other. The Zone Read is the method of determining which option to use. The idea behind it is to create an advantage in terms of the number of players on the offense surrounding the football compared to defense. Although both teams have 11 men on the field, because the Quarterback is not used to block, when he is not carrying the ball, the defense has an extra player to use against the offense in the blocking scheme. In the Zone Read, the Offensive Line allows this extra player to move freely, but then chooses the option that puts the ball in the opposite direction of where he is moving.[7]

2.3.3 Defensive Terminology

1. **3-4 Defense:** This is a defensive formation that puts three men on the Defensive Line and four Linebackers behind them. The Linemen involved in this scheme are two Defensive Ends and one Nose Tackle in the middle. The Linebackers are two Outside Linebackers and two Inside Linebackers.

2. **4-3 Defense:** This is a defensive formation that puts four men on the Defensive Line and three Linebackers behind them. It is the most common and popular defensive formation used. The Linemen involved in this scheme are two Defensive Ends and two Defensive Tackles. The Linebackers are two Outside Linebackers and one Middle Linebacker.

3. **Blitz:** This is when the defense charges directly for (the passer) as soon as the ball is snapped (it is also referred to as red dogging). In order to execute this play and get through the defense to put pressure on the Quarterback, the defense will rush more football players than the offense has accounted for. From an offensive point of view, this means there are more men for them to block than usual and it crowds the space around the quarterback, which can lead to a sack. However, the risk the defense takes is that while they are occupied with getting to the Quarterback, the additional defensive player has left his normal position and this leaves more space for the Receiver to get open to catch the ball or to run in once he has caught the ball. The Linebackers are the defensive players usually involved in a Blitz, but it can also be the Defensive Backs (Safeties and Corner-backs). Where they are positioned before the snap can be misleading as some players only "show blitz" when they are bluffing.

4. **The Box:** This is the area on the defensive side of the field in the middle and directly behind the Line of Scrimmage. It encompasses the area from the Defensive Line to approximately five yards back, typically containing the Linebackers, and its outer edges are determined by where the Defensive Ends line up on either side. The standard formation puts 7 Men in the Box. There are different defensive strategies relating to the Box such as "8 Men in the Box" which is a defensive formation to put additional concentration on stopping the run and "6 Men in the Box" which is a defensive formation that takes players out of the Box in order to put additional pressure on the passing game.[6]

5. **Bump-and-Run:** A technique used bypass defenders (Defensive Backs) in which they initially hit the receiver "bump" within one yard of the line of scrim-

mage (in the NFL this number is five yards). They are not allowed to touch him past that range until he actually has the ball. Then they follow him or "run" with him. This slows down the offensive player in an attempt to throw off the timing of the route he is running and when and where the quarterback expects him to be. And because he has been slowed down it also gives the defensive player time to run with him and be in a position to prevent him from catching a pass.

6. **Close the Cushion:** Whatever space is between the offensive player and himself, the defensive player must recover this or close the cushion once the ball is in the air. Read more about this under "Playing Defensive Back." Control the Cushion: This means that the defensive back is limiting the cushion or distance between the receiver and himself all while the receiver is trying to expand that space. If you can't touch the receiver you can't defend the pass. The distance between them is the battle of who controls the cushion and you can tell who is winning by how close or far apart they are. Read more about this under

7. **Cushion:** This is the space between the defensive player (usually the defensive back) and the receiver. The defensive back tries to maintain a cushion of three to four yards between him and the receiver so that he can touch him. The receiver, on the other hand, will try to get separation from the defensive player (i.e., a bigger cushion). Read more about this under "Playing Defensive Back." Double Coverage: Also known as Double Teaming. This is when two defensive players are covering one offensive player.[8]

8. **Gap:** Any space on the Defensive Line of Scrimmage that isn't physically occupied by a body is called a Gap. A Defensive Linemen has two Gaps on either side of him and will be responsible for either one or both of them. In certain defensive formations it is important to have 2-Gap Defensive Tackles who, by defending both Gaps, thereby take on two players at once and free up an additional defensive player to protect against the pass.

9. **Man-to-Man:** This is a type of defensive scheme. There are two basic defen-

sive schemes and they are distinguished based on how they defend against the pass. In Man-To-Man, as the name suggests, the pass defenders are assigned to defend against specific players.

10. **Pass Rush:** This is used to describe what the defensive players are doing when they attempt to get through the offensive players and to the player (usually the Quarterback) attempting to pass the ball. If they get to the quarterback while he still has the ball in his hand, it is considered a sack.

11. **Prevent Defense:** This is when four or more players are assigned to cover the deep passing threat. With less players up front, the defense sacrifices the run and short pass to avoid giving up big plays. They allow this to happen because the hope is that the clock will expire before the other team can make it far enough down the field to score. However, because the Quarterback is often able to make short and medium length passes, the offense will eventually move the ball down the field. Some say that that the only thing the Prevent Defense does is prevent you from winning.[6]

12. **Zone Defense:** This is a type of defensive scheme. There are two basic defensive schemes and they are distinguished based how they defend against the pass. In Zone Coverage, the pass defenders cover a specific area or zone on the field.[7]

Chapter 3 Data Exploration

Here, we first introduce the reader to some basic notions of statistical analysis. Then, we present the data followed by various basic statistics and plots of the variables from the data.

3.1 Basic Statistical Notions and Methods

1. **Level of significance:** Also known as *alpha level*. This value is used as a probability cutoff for making decisions about the null hypothesis. It value represents the probability we are willing to place on our test for making an incorrect decision in regards to rejecting the null hypothesis. In other words, it is the level of risk we are willing to take as we reject a possibly correct hypothesis. For example, a significance level of 0.05 indicates a 5% risk of concluding there is a statistically significant result or difference when there is none.

2. **P-value and Confidence Interval:** P-values (labelled *Sig.*, in SPSS) are the probability of obtaining an effect or a relationship at least as extreme as the one in the sample data, as we assume the truth of the null hypothesis. When a p-value is less than or equal to the significance level (typically 0.05), we reject the null hypothesis.

   The range of values, for which the p-value exceeds a specified alpha level is called **confidence interval**. In other words, this interval gives a range of values within which lies a true (population) parameter. So, with an estimated parameter at $\alpha = 0.05$, a confidence interval indicates that, with repeated samplings (identical studies in all respects except for random error), we are "confident" that, in spite of margin-of-error (or deviations), 95% of the parameter estimates will lie within

this interval. With the margin-of-error we can state that the interval includes the true population parameter.

3. **Correlation:** A simple correlation measures the relationship between two (ideally normally distributed) variables. For our thesis we used Pearson's r which measures a linear relationship (or association) between two continuous (numeric) variables without taking into account other variables. For each pair of variables (X_i, X_j) Pearson's correlation coefficient is computed using

$$r = \frac{\sum_{i=1}^{n}(x - \overline{x_i})(y - \overline{y_i})}{\sqrt{\sum_{i=1}^{n}(x - \overline{x_i})^2 \sum_{i=1}^{n}(y - \overline{y_i})^2}}.$$

Its value range between -1 and 1 and $|r| \sim 1$ indicates a strong dependence or correlation and $|r| \sim 0$ indicates a strong independence between the variables.

The objective of any data analysis is to extract information (or accurate estimation) from the original (raw) data. Typically, we seek to determine whether or not there is statistical relationship between a response variable (Y) and explanatory variables (X_i). One way to answer this question is to use some regression analysis in order to *model* its relationship. By modeling we try to predict the outcome (Y) based on values of a set of predictor variables (X_i). There are several types of regression analysis and each type of the regression model depends on the type of the distribution of Y. They are often used to assess the impact of multiple variables (a.k.a. co-variates and factors) in the same model. Here, we focus on two of these which we define next.

4. **Linear regression:**

This is an extension of the simple correlation. In regression, one or more variables X_i (*predictors* or *factors* or *independent variables* or *inputs*) are used to predict an outcome Y_i (*response* or *target* or *criterion* or *dependent variable* or *output*). In practice, a linear regression model or equation returns estimates of the coefficients of a linear equation that involves one or more independent variables that best predict the values of an output or the dependent variable

which must be quantitative continuous or scale. It is often written as

$$E(Y_i) = \beta_0 + \beta X_i \text{ or } Y_i = \beta_0 + \beta X_i + \epsilon_i$$

for each i observation or data point with errors ϵ_i.

Regression coefficients or coefficient estimates β_i represent the mean change in the response variable for one unit of change in the predictor variable while holding other predictors in the model constant.

The *p-value* for each term tests the null hypothesis that the coefficient is equal to zero (no effect). Thus, a low p-value (< 0.05) indicates that we can reject the null hypothesis, in which case the corresponding predictor is likely to be a meaningful addition (or is *statistically significant*) to your model. Likewise, a larger (insignificant) p-value suggests that changes in the predictor are not associated with (or do not help explain) changes in the response. Thus, for our analysis, we use the coefficient p-values to determine which variables are useful for our final model.

As it is true for any model, part of the process involves checking to make sure that the data we want to analyze can actually be done using the chosen model. For a linear model it is required that, for each value of the independent variable, the distribution of the dependent variable must be normal. Typically, we plot the errors (residuals) to see if they follow a normal distribution. A QQ-plot is an example of such a residual plot that can be used to reveal biased results more effectively than a simple computation. Further, the variance of the distribution of the dependent variable should be constant for all values of the independent variable. Finally, the relationship between the dependent variable and the independent variables should be linear, and all observations should be independent. In brief, the residuals of a good model should be normally and randomly distributed.

In the event the response variable takes a form where the residuals look completely different from a normal distribution, it is preferable to consider another

class of models known as *generalized linear models (GLM)*; in which case the response variable Y_i follows an exponential family distribution. Logistic regression is an example of a GLM as we define it, next.

5. **Binomial Logistic regression:**

   Binomial Logistic regression which is simply called a *logistic regression* estimates the probability of an occurrence of an event Y_i based on a set of predictors X_i. The basic mathematical concept behind logistic regression is *logit* which is the natural logarithm (ln) of an odds; and odds are ratios of probability "success" p (for instance, an ambulance was needed) to probability "failure" $1-p$ (when no ambulance was needed, for instance). Thus, given a response categorical variable Y and m predictors X_i, we have

$$logit(Y) = log(\frac{p}{1-p}) = \beta_0 + \sum_{i=1}^{m} \beta_i X_i \tag{3.1}$$

   where β_0 is the Y intercept (i.e., mean of Y independent of X_i's) and β_i's are the *regression coefficients* (or *parameter estimates*) for each predictor X_i, for $i = 1, \ldots, m$.

   We note that, by taking exponential (or antilog) of both sides of equation 3.1, we derive the equation to predict the probability of the occurrence of an outcome of interest as follows:

$$\begin{aligned} p &= Probability\ (Y = outcome\ of\ interest \mid X_1 = x_1,\ X_2 = x_2,\ \ldots,\ X_m = x_m) \\ &= \frac{e^{\beta_0 + \sum_{i=1}^{m} \beta_i X_i}}{1 + e^{\beta_0 + \sum_{i=1}^{m} \beta_i X_i}}, \end{aligned}$$

   where $e \sim 2.71828$ is the natural base.

   Interpretation :

   (i) The sign $(\pm)$ of a coefficient (or slope) β_j gives the direction of the relationship (negative or positive) between the predictor X_j's and the logit of Y.

   (ii) The intercept or log average odd $\beta_0 = log(\frac{p}{1-p})$ is an estimate of the model

(null model) if we consider no predictor; this is also known as *unconditional log odds* of the response. Thus, the **average odd** is e^{β_0} and the **average probability of success**, p is $\frac{e^{\beta_0}}{1+e^{\beta_0}}$.

(*iii*) The coefficient β_j, for some predictor X_j. Fixing the levels of the remaining predictors X_k, $k \neq j$, this value gives the log(odds) of the effect of X_j on Y (beyond the average) for each unit increase (in a scale variable) or in comparison to a fixed (base) level in X_j. Thus, for a predictor X_j, the **estimated odds** value is e^{β_j} and the **percentage change** in odds (per unit increase or relative to a base level) is

$$(e^{\beta_j} - 1) \times 100\%.$$

As related to inferential statistics, a *null hypothesis* would state that, for some $\beta_j = 0$, $j > 0$, i.e., there is no linear relationship between logit of Y and X_j, in the population. So, rejecting such a null hypothesis would imply that a linear relationship exists between logit of Y and X_j. As indicated earlier for linear regression, we will rely on the p-values and the alpha level of .05, to help make our decision on the significance of the coefficients.

6. **Multi-nomial Logistic regression:**

   Multinomial logistic regression (or *multinomial regression*) is used to predict a nominal dependent variable (with two or more factors or categories) given one or more independent variables. As such, it is an extension of binomial logistic regression to allow for a dependent variable with more than two categories.

7. **R-squared:**

   Also known as *coefficient of determination.* it is a statistical measure of how close the data are to the fitted regression line. In other words, it is the percentage of the response variable variation that is explained by a linear model in which case

$$R^2 = \frac{Explained\ variation}{Total\ variation} \times 100$$

   0% indicates that the model explains none of the variability of the response data

around its mean and 100% indicates that the model explains all the variability of the response data around its mean. In general, the higher the R-squared, the better the model fits your data but there are risks of "overfitting" or bias, which makes the model less adaptable to a different data taken under a similar circumstance.

8. **Pseudo R-squared:**

   As opposed to an R-squared value that is obtain from evaluating a model built on a continuous response, such an indicator does not make sense for models built on an ordinal response where the variance is fixed instead. However, a similar metric (in scale) called a "Pseudo" R-squared is used for models such as logistic regressions. In which case, the higher the value the better model but they are only meaningful when comparing these values for distinct models. There are several such pseudo R-squared values but SPSS software returns the values for Nagelkerke, and Cox & Snell (Pseudo) R-squareds.

   There are other types of analysis which might be of interest for further readings. However, our thesis does not include any of them. We list them as next.

9. **Factor analysis:**

   Is a regression based data analysis technique, used to find an underlying structure in a set of variables. It goes with finding new independent factors (variables) that describe the patterns and models of relationships among original dependent variables. Factor analysis is a very popular tool for researching variable relationships for complex topics such as psychological scales and socioeconomic status. Factor Analysis is a basic step towards effective clustering and classification procedures.

10. **Dispersion analysis:**

    Dispersion analysis is not a so common method used in data mining but still has a role there. Dispersion is the spread to which a set of data is stretched. It is a technique of describing how a set of data sets extended. Generally, the

dispersion has two matters: first it represents the variation of the things among themselves, and second, it represents the variation around the average value. If the difference between the value and average is significant, then dispersion is high. Otherwise, it is low.

11. **Discriminant analysis:**

    Is one of the most powerful classification techniques in data mining. The discriminant analysis utilizes variable measurements on different groups of items to underline points that distinguish the groups. These measurements are used to classify new items. Typical examples of this method uses are: in classifying applications for credit cards into low risk and high-risk categories, classifying customers of new products into different groups, medical studies implicating alcoholics and non-alcoholics.

12. **Time series data analysis:**

    Is the process of modeling and explaining time-dependent series of data points. The goal is to draw all meaningful information (statistics, rules, and patterns) from the shape of data.

13. **Evolutionary Programming:** Evolutionary programming in data mining is a common concept that combines many different types of data analysis using evolutionary algorithms. Most popular of them are: genetic algorithms, genetic programming, and co-evolutionary algorithms. In fact, many data management agencies apply evolutionary algorithms to deal with some of the world's biggest big-data challenges.

3.2 Variables

The regular season began on August 24, 2019, and ended on December 14, 2019. The postseason concluded on January 13, 2020, with the 2020 College Football Playoff National Championship at the Mercedes-Benz Superdome in New Orleans. This was the sixth season of the College Football Playoff (CFP) championship system. The

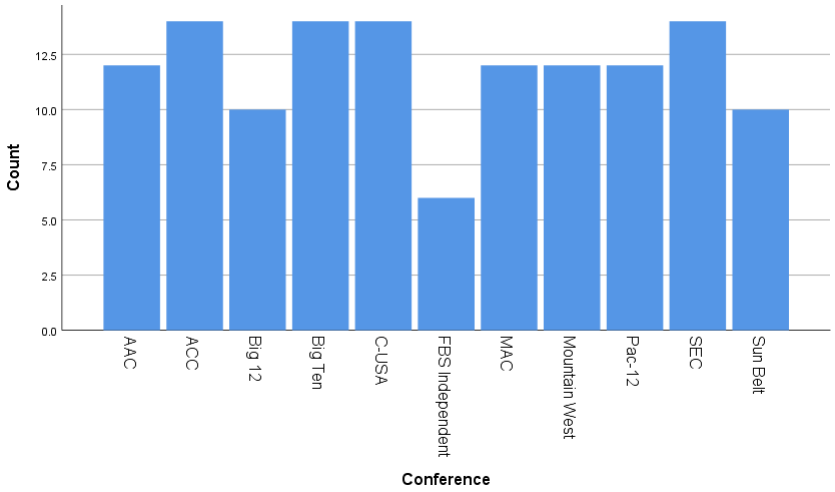
original data for the season is an excel spreadsheet generated for the 2019 College football team performance/stat, extracted from `https://www.teamrankings.com/ncf/team-stats/`. Then, we visit each of the following thee tabs: Scoring Offense, Total Defense and Turnovers. From each three tabs, we pick the desired variable, copy-paste, the information for all the teams into an excel file. After editing the file (mostly removing unwanted columns, matching records and merging files), the resulting file contains 131 rows (1 header, 130 team names) with 10 columns which are Conference, WinPercent, WinMargPer, PtPerGame, PtPerPlay, PtPerPlayMargin, OffPtPerGame, Opp3rdownConv, OppPuntPerPlay, TurnMarg. All these variables are numerical (scale in SPSS) except Conference which is categorical (nominal in SPSS). Here, we explore some of the variables.

3.2.1 Conference

There are a total 11 conferences with 130 teams as shown in the frequency table. SEC, ACC, Big ten and C-USA have the most team members (14) while FBS Independent has the least team memberships (6).

Conference

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	AAC	12	9.2	9.2	9.2
	ACC	14	10.8	10.8	20.0
	Big 12	10	7.7	7.7	27.7
	Big Ten	14	10.8	10.8	38.5
	C-USA	14	10.8	10.8	49.2
	FBS Independent	6	4.6	4.6	53.8
	MAC	12	9.2	9.2	63.1
	Mountain West	12	9.2	9.2	72.3
	Pac-12	12	9.2	9.2	81.5
	SEC	14	10.8	10.8	92.3
	Sun Belt	10	7.7	7.7	100.0
	Total	130	100.0	100.0	



3.2.2 Winning Percentage

The winning percentage appear to be relatively normal as shown in Figure 3.1. It is computed as

$$Win\ percentage = \frac{\#Win}{\#Game} \times 100\%$$

The average Win is 53%. The Team with the most win is LSU (SEC) with 100% and the team with least win is Akron (MAC) with 0%.

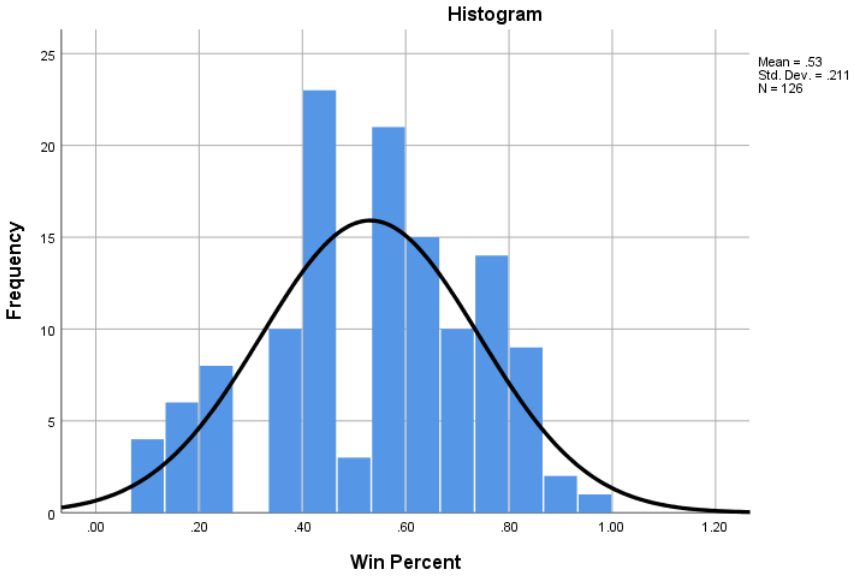


Figure 3.1: 2019 Win Percentage Distribution

3.2.3 Winning Margin Percentage

Because different teams have different total number of games, we compute each team winning margin as:

$$Win\ Margin\ percentage = \frac{(\#Win - \#Loss)}{\#Game} \times 100\%$$

We test the relation between Win Percent and Win Margin. Naturally, we find a strong correlation between these two variables as shown in Figure 3.2. So, we decided not to keep them both in any model discussed in Chapter 5.

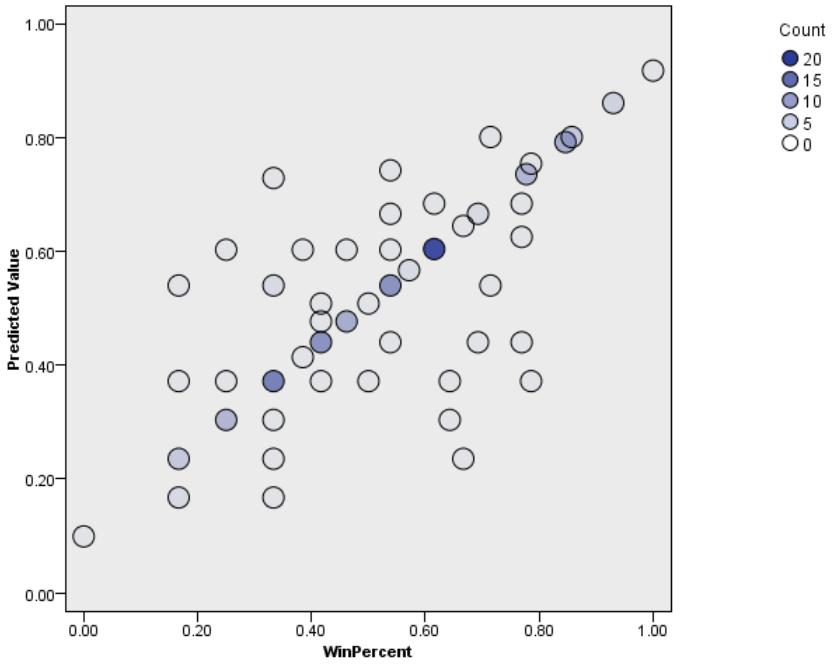


Figure 3.2: Scatter Plot of Win Percent vs Win Margin Percent

3.2.4 Binned Winning Margin

Because there are 11 college conferences reported in our data, we decided to bin the Win Margin variable to see whether it has a relation with Conference. In other words, do teams in certain conferences have larger Win Margin than others. Figure 3.3 appears to suggest this. For instance, SEC and Big Ten conferences show the teams with largest positive Win Margins while MAC and Sun Belt conferences have the teams with largest negative (or loss) Win Margins.

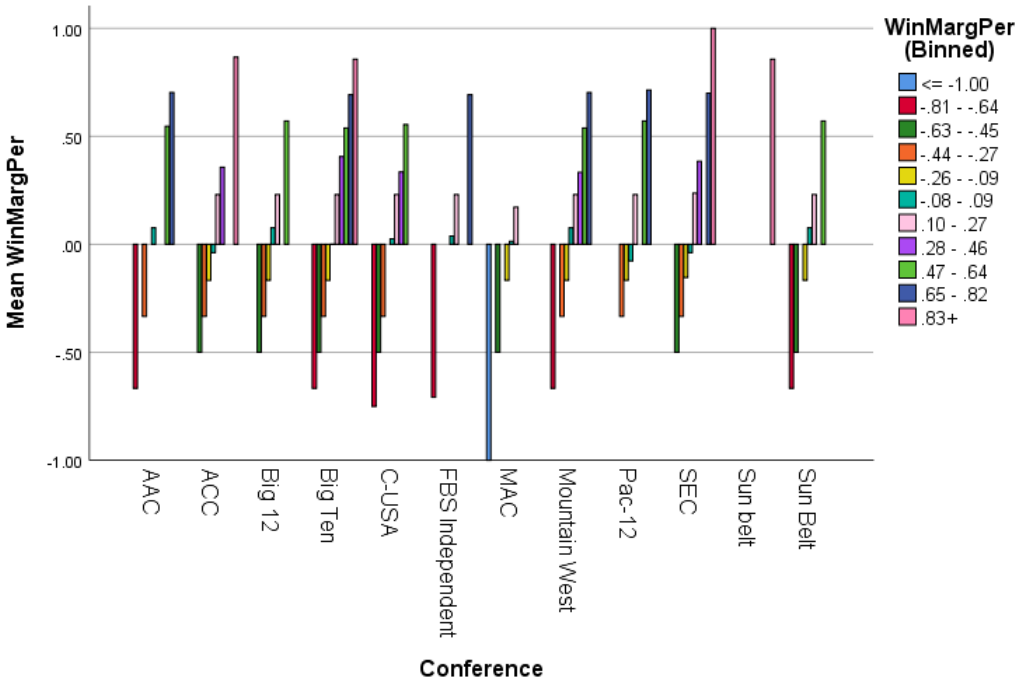


Figure 3.3: Histogram of Win Margin Within each Conference

This previous observation is also supported by Figure 3.4 where one team in MAC (Akron) shows an unusually low Win Margin value. We also note that an ACC team (Clemson) is an outlier for this conference for their Win Margin. Likewise, a Sun Belt team (LA Lafayette) is an outlier for this conference for their Win Margin.

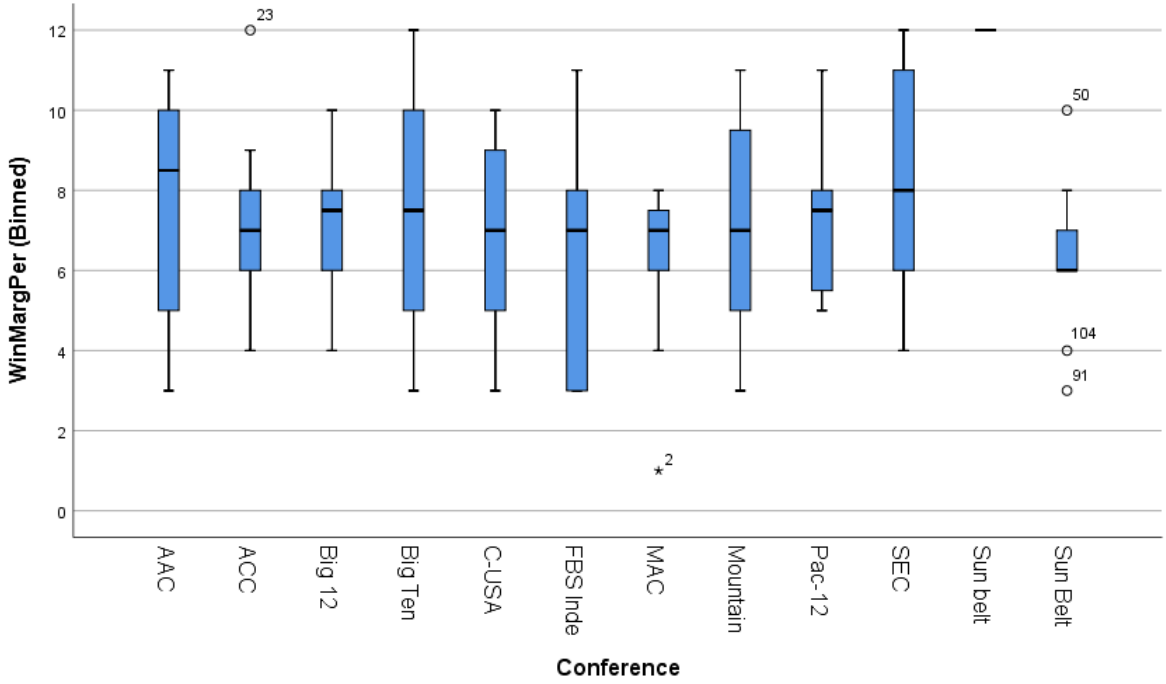


Figure 3.4: BoxPlot of Win Margin by Conference

3.2.5 Point Per Game

Figure 3.5 shows that the minimum average point per game is 10.5 and the maximum is about 50 points. The mean is 28 points per game with a standard deviation of 7 points per game. The point per game throughout the 2019 season is normally distributed.

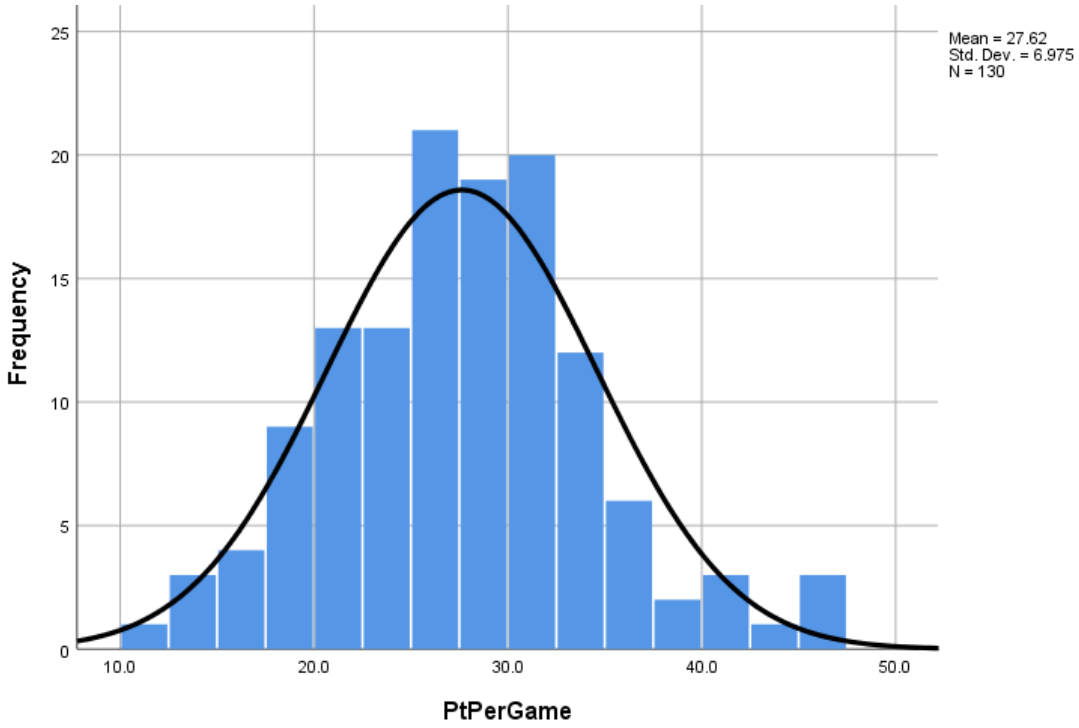


Figure 3.5: Offensive Plays vs Frequency

3.2.6 Point Per Play

Figure 3.6 shows that the minimum average point per play is 2 and the maximum is 7 points. The mean is about 4 points per game with a standard deviation of 1 point per game. The point per play throughout the 2019 season is normally distributed.

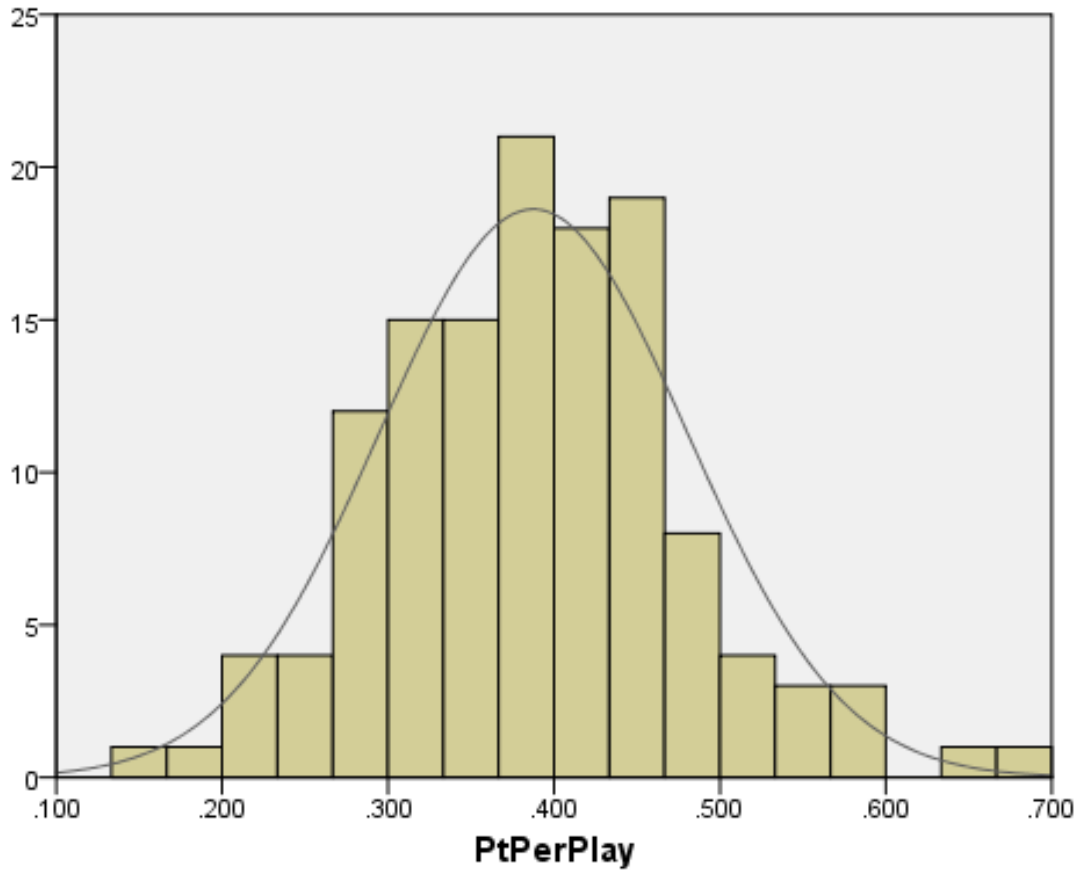


Figure 3.6: Team Point per Play distribution

3.2.7 Offensive Play

The average offensive plays for the 2019 season is about 886 with the lowest number of plays being 717 (team: Georgia Tech) and the maximum number of plays being 1080 (team: LSU) The figure below is showing the number of Offensive Plays ran by a team out of the 130 teams used in this data collection represent as N=130. An average team runs around 800 to 1000 plays within one season.

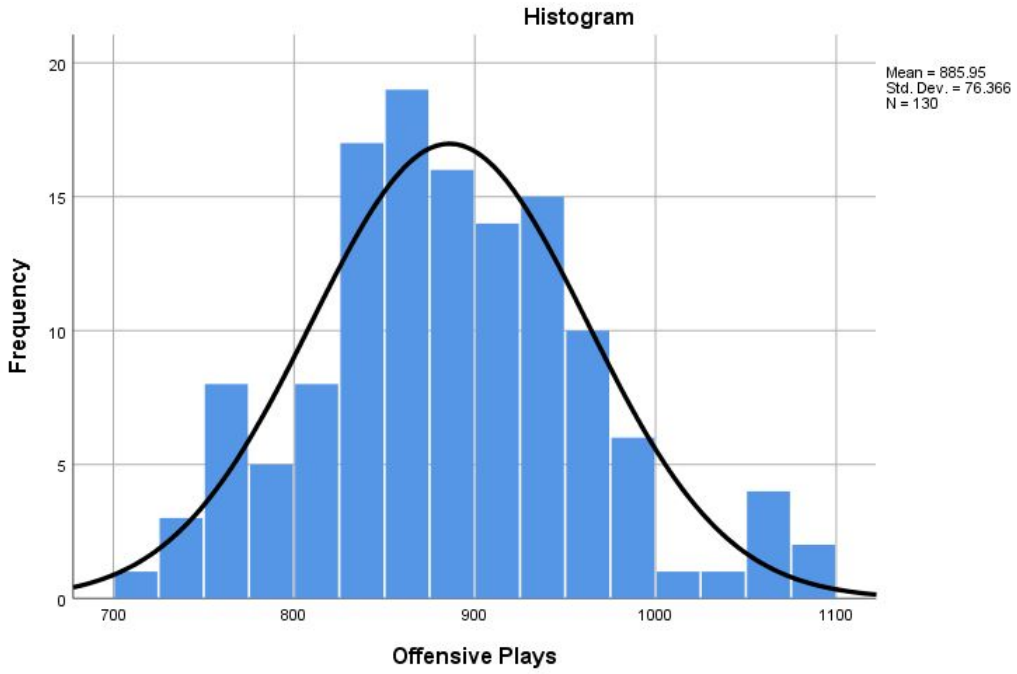


Figure 3.7: Frequency of Offensive Plays

3.2.8 Other Variables

Other variables statistics are shown in the Figure 3.8.

		PtPerPlayMargin	OffPtPerGame	Opp3rdownConv	TurnMarg
N	Valid	130	130	130	130
	Missing	0	0	0	0
Mean		-.00745	26.481	.3978	-.014
Std. Deviation		.160178	6.6640	.05796	.6164
Minimum		-.451	9.9	.27	-1.5
Maximum		.407	46.3	.54	1.6

Figure 3.8: Summary Statistics of other variables

Chapter 4 Data Analysis

Here, we rely on SPSS (Statistical Package for Social Sciences) which is a statistical software offered by IBM for complete data analysis; it is a family of advanced computer programs for statistical analysis. We first run a linear regression on the continuous variable Win Margin, and then, we run a logistic regression on Win Margin, after we bin it as a binary variable. A team with a Win Margin $> 50\%$ is likely to be qualified for a Bowl game–So, if Win Margin $> 50\%$, Win Margin is assigned a value of 1 and 0, otherwise.

4.1 Multivariate Linear Model

As mentioned in Chapter 4 data exploration, there is strong correlation between Win Margin Percentage and Win Percentage so for our model analysis we use Win Margin and later Binned Win Margin as our response variables. As a selection criterion, we use *Akaike information criterion (AIC)* which is an estimator of out-of-sample prediction error; it is therefore a comparative tool (step-wise selection) designed to estimate the quality of a model, relative to other models. Negative AIC indicates less information loss than a positive AIC and therefore a better model; so the small the AIC value the better the model. Another model performance measure (for linear regression) is R^2 which we defined in Chapter 3.

4.1.1 Variable Selection Process

We apply a forward step-wise selection (AIC criterion) to determine most influential predictor for Win Margin. Figure 4.2 shows that Point Per Play Margin along with Opponent 3rd Conversion and Point Per Game have the biggest effect. The AIC

value is -424.4 with an Adjusted R^2 value of .8. Further, among these variables, the Opponent 3rd down conversion has negative effect on Win Margin while other predictors have positive effects as shown in the Figure.

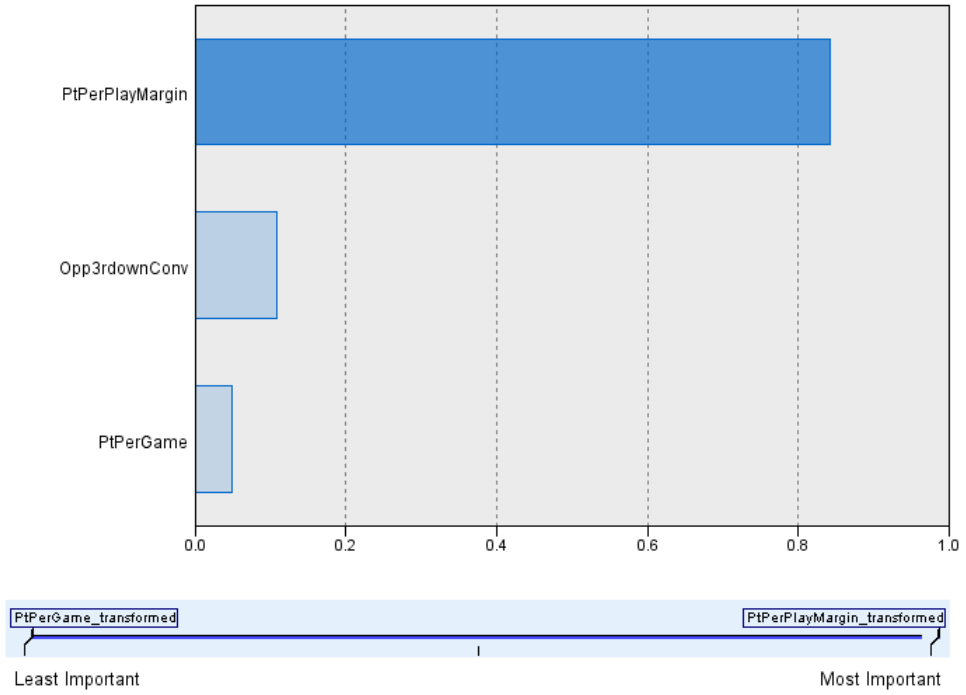


Figure 4.1: Variable Importance on Win Margin

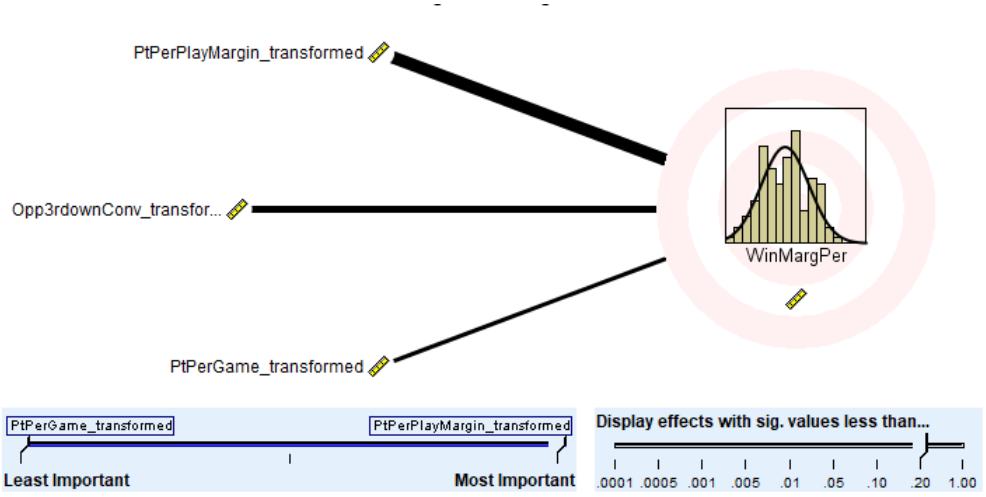


Figure 4.2: Predictors Effects on the Response Win Margin

Such linear regression appears to meet a basic linear regression model assumption which is a normal distribution of the residuals as shown in Figure 4.3.

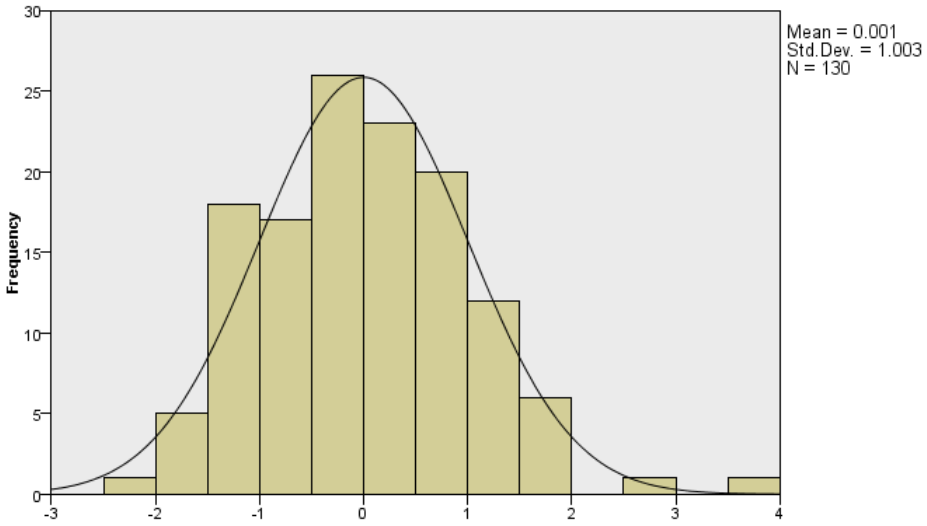


Figure 4.3: Studentized Residual Distribution

Moreover, the predicted values by the model appear to correlate with the observed value for the win margin as shown in Figure 4.4.

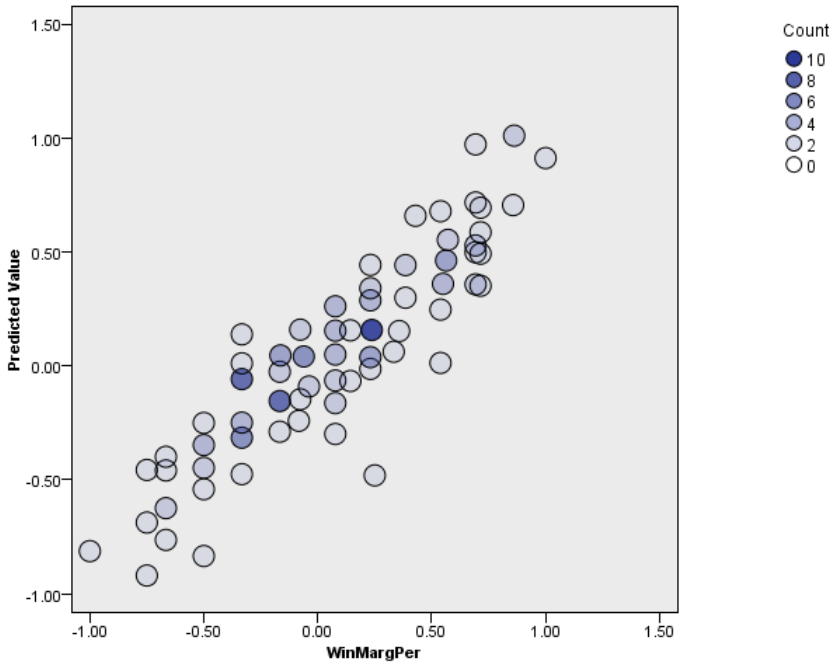


Figure 4.4: Predicted by Observed

We anticipate a similar analysis out-put for Win Percent variable, so we choose not to repeat the previous analysis for this variable.

4.1.2 Enhance Model Accuracy (Boosting)

We decide to enhance the model (Gradient boosting) and compare the performance of the new (Ensemble) model to the the original (Reference Model). We found that two additional variables, Offensive Point Per Game and Opponent Punt Per Play show some effect on the model (see Figure 4.5), although minimal. They do not improve the model significantly as shown in Figure 4.6). Hence, adding more variables does not necessarily improve the model significantly.

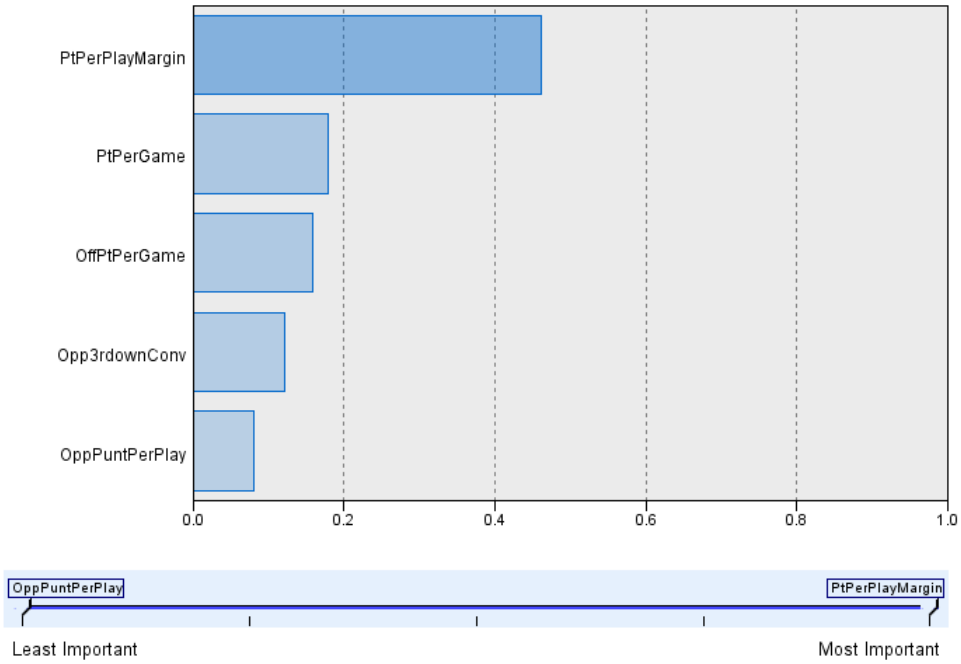


Figure 4.5: Predictors importance for Enhanced Model

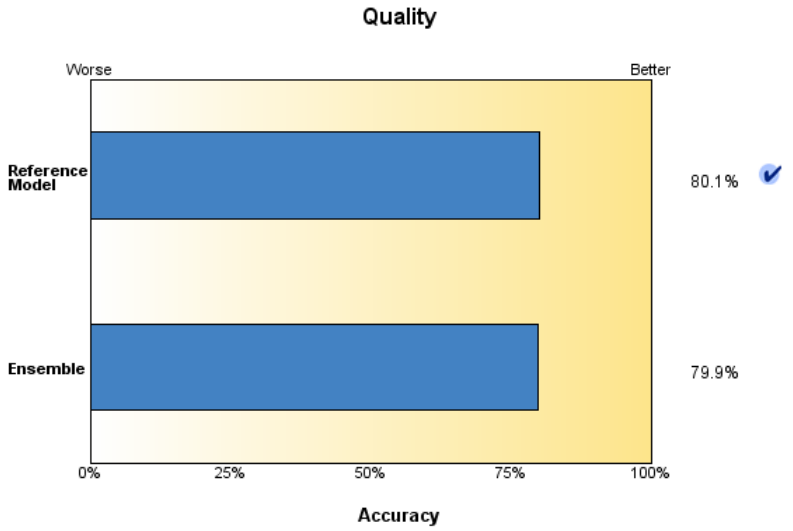


Figure 4.6: Models Performance Summary

4.1.3 Linear Regression Results

The result on a linear regression on Win Margin, on all 7 predictors is shown in Figure 4.7.

Model		Unstandardized Coefficients		Standardized Coefficients			95.0% Confidence Interval for B	
		B	Std. Error	Beta	t	Sig.	Lower Bound	Upper Bound
1	(Constant)	.372	.388		.957	.341	-.397	1.140
	PtPerGame	.026	.024	.433	1.115	.267	-.021	.073
	PtPerPlay	-.226	.973	-.049	-.232	.817	-2.153	1.701
	PtPerPlayMargin	1.758	.323	.661	5.439	.000	1.118	2.398
	OffPtPerGame	-.017	.019	-.260	-.860	.392	-.055	.022
	Opp3rdownConv	-1.170	.538	-.159	-2.174	.032	-2.236	-.105
	OppPuntPerPlay	-.649	2.247	-.021	-.289	.773	-5.097	3.799
	TurnMarg	.027	.038	.038	.701	.485	-.048	.101

Figure 4.7: Regression Coefficients

1. **Observation**

   The predictors Point Per Play Margin (with $p \sim .000$) and Opponent 3rd Down Conversion (with $p < .05$) are the two variables that are significant.

2. **Regression Equation**

$$Y = .4 + .1.8X_1 - 1.2X_2,$$

   where

   Y:= The average predicted Win Margin

   X_1:=Point Per Play Margin

   X_2:=Opponent 3rd Down Conversion

3. **Interpretations**

   (i) For each 1 additional score or point by a team, their Win Margin increases by 2.2 above the average Win Margin.

   (ii) For each 1 additional 3rd Down Conversion of the opposing team, the Win Margin decreases by 1.2 below the average Win Margin.

   (iii) We are 95% confident that the true estimate for the coefficient of X_1 is within $[1.1, 2.4]$ while the true estimate for for the coefficient of X_2 is within

$[-2.24, -1.11]$.

4. **Model Summary**

Figure 4.8 indicates that we have a robust predictive model with $R^2 \sim .9$. In other words, about 90% of the variations in Win Margin can be explained by the two significant predictors, namely Point Per Play Margin and Opponent 3rd Down Conversion.

					Change Statistics				
Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	R Square Change	F Change	df1	df2	Sig. F Change
1	.897[a]	.804	.792	.19420	.804	71.372	7	122	.000

Figure 4.8: Regression Coefficients

5. **Model Assumptions**

The residual plot and the normal $P-P$ plot of the regression indicate the model has met the basic linear regression model assumptions of normality, randomness of error. See Figures 4.9 and 4.10.

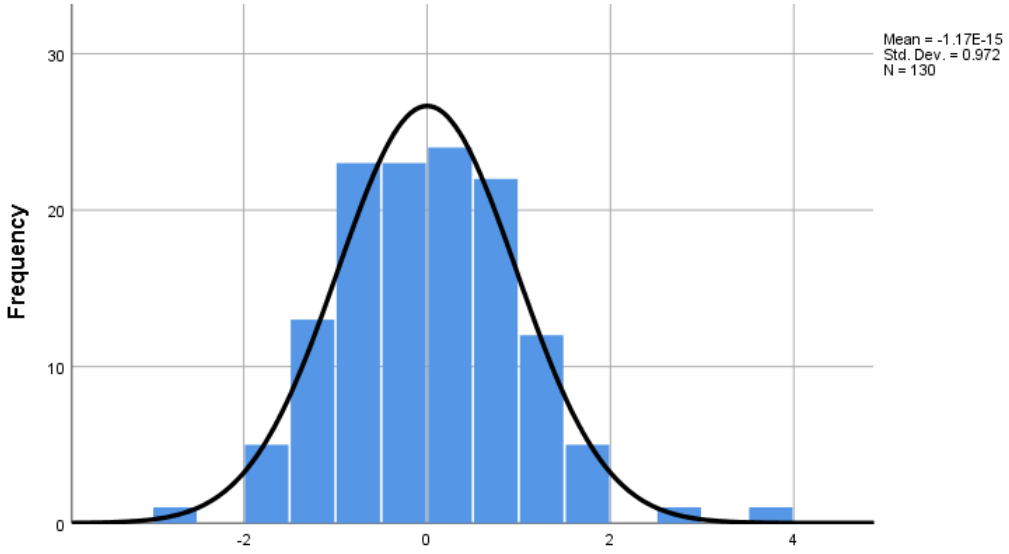


Figure 4.9: Regression Standardized Residual Plot

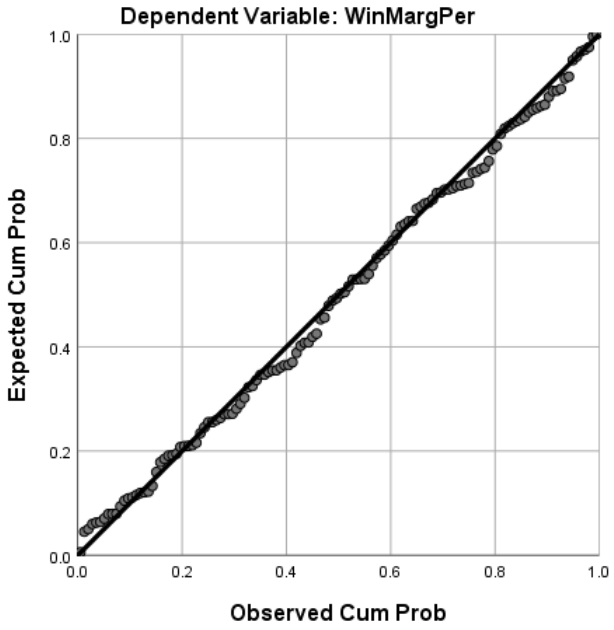


Figure 4.10: Normal P-P plot of Regression Standardized Residual Plot

4.2 Multivariate Logistic Regression

For this reason, we run a multiple logistic regression on Conference with predictor (binned) Win Margin. We found that, according the model, it is very hard to classify teams based on Win Margin. Meaning, Win Margin does not determine or explain Conference affiliation. So, as shown in Figure 4.11, only AAC conference teams have an acceptable classification rate (about 67% accuracy) while the remaining conference teams are poorly (with accuracy $\leq 50\%$) classified.

	Predicted												
Observed	AAC	ACC	Big 12	Big Ten	C-USA	FBS Independent	MAC	Mountain West	Pac-12	SEC	Sun belt	Sun Belt	Percent Correct
AAC	8	0	0	0	0	1	1	0	0	2	0	0	66.7%
ACC	1	4	0	3	0	0	4	0	0	0	0	2	28.6%
Big 12	3	3	0	1	0	0	1	0	0	0	0	2	0.0%
Big Ten	2	1	0	5	0	1	2	0	0	2	0	1	35.7%
C-USA	5	1	0	3	0	2	3	0	0	0	0	0	0.0%
FBS Independent	0	1	0	0	0	2	2	0	0	1	0	0	33.3%
MAC	0	3	0	1	0	0	6	0	0	0	0	2	50.0%
Mountain West	4	1	0	1	0	1	2	0	0	2	0	1	0.0%
Pac-12	4	4	0	0	0	0	1	0	0	1	0	2	0.0%
SEC	2	3	0	3	0	0	2	0	0	3	0	1	21.4%
Sun belt	0	0	0	1	0	0	0	0	0	0	0	0	0.0%
Sun Belt	1	1	0	1	0	1	2	0	0	0	0	3	33.3%
Overall Percentage	23.1%	16.9%	0.0%	14.6%	0.0%	6.2%	20.0%	0.0%	0.0%	8.5%	0.0%	10.8%	23.8%

Figure 4.11: Team Conference Classification by Win Margin

Futher, we bin the Win Margin as a binary variable. A Win Margin of 50% or less is coded as 0; there are 60 teams associated with such value. A Win Margin value greater than 50% is coded as 1 and there are 70 such teams. We found (see Figure

4.12) that the new model has an accuracy rate of 86% in separating teams with a Win Margin of 50% or less vs teams with a Win Margin greater than 50%. Figure 4.13 shows the model ouput as we consider all predictors in the model.

			Predicted		
			WinMargPer (Binned)		Percentage Correct
	Observed		0	1	
Step 1	WinMargPer (Binned)	0	51	9	85.0
		1	9	61	87.1
	Overall Percentage				86.2

Figure 4.12: Team Classification on Win Margin at a .5 cut value

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	Conference			5.336	11	.914	
	Conference(1)	-1.356	1.544	.772	1	.380	.258
	Conference(2)	-1.794	1.463	1.503	1	.220	.166
	Conference(3)	-2.245	1.643	1.866	1	.172	.106
	Conference(4)	-2.840	1.655	2.944	1	.086	.058
	Conference(5)	-1.285	1.515	.720	1	.396	.277
	Conference(6)	-.159	1.883	.007	1	.933	.853
	Conference(7)	-.839	1.409	.355	1	.552	.432
	Conference(8)	-.465	1.391	.112	1	.738	.628
	Conference(9)	-1.318	1.578	.697	1	.404	.268
	Conference(10)	-2.534	1.700	2.221	1	.136	.079
	Conference(11)	11.902	40192.970	.000	1	1.000	147626.578
	PtPerPlayMargin	16.461	6.884	5.718	1	.017	14089520.50
	PtPerGame	.034	.434	.006	1	.937	1.035
	PtPerPlay	2.335	18.273	.016	1	.898	10.334
	OffPtPerGame	-.025	.349	.005	1	.942	.975
	Opp3rdownConv	-14.940	11.126	1.803	1	.179	.000
	OppPuntPerPlay	44.393	51.581	.741	1	.389	1.904E+19
	TurnMarg	.123	.699	.031	1	.860	1.131
	Constant	3.883	8.909	.190	1	.663	48.589

a. Variable(s) entered on step 1: Conference, PtPerPlayMargin, PtPerGame, PtPerPlay, OffPtPerGame, Opp3rdownConv, OppPuntPerPlay, TurnMarg.

Figure 4.13: Binary logistic regression output on Win Margin

Model Equation:

$$\ln(y) = .4 + 16.5x_1$$

where,

y_1 :=*Predicted odds of having a Win Margin* $> 50\%$

x_1:=Point Per Play Margin

Interpretations:

The log odds of ending the 2019 football season with a Win Margin greater than 50% increases by 16.5 above the log average odd for each point per play margin.

Chapter 5 Conclusion and Future Research

Predicting team win or rank is hard, as many factors play some roles. In this research we found that some parameters such as Point Per Play Margin have consistently remained influential whether we consider Win Margin as continuous for a linear regression analysis or as binary for a logistics regression analysis. This should not come as a surprise as a deterministic factor in team wining a game, hence a season. We also found that Conference has no effect on winning percentage. Some teams such as Akron (MAC) had lost all games and LSU (SEC) had won all games and was the championship winner for the 2019 NCAA football season.

We run into some issues related to getting the right data that includes essential predictors. The most complete data we obtained came from the 2019 season, hence the focus of this research. Still, we have managed to compile some data for the previous three years (2018-2016) that include the same predictors, and yet, we found some errors or discrepancies in the last minutes. Future research can include some past years data to obtain a more robust predictive model.

Bibliography

[1] Long, Joseph. Development of a Prediction Model for the NCAA Division-I Football Championship Subdivision. Diss. North Dakota State University, 2013.

[2] Berg, Kris, Richard W. Latin, and Thomas Baechle. "Physical and performance characteristics of NCAA Division I football players." Research quarterly for exercise and sport 61.4 (1990): 395-401.

[3] Harris,Christopher M.,and Gary C. McMahan. "Human capital stability: the influence of overlapping tenure on the performance of NCAA football teams." American Journal of Management 13.3 (2013): 78-93.

[4] Spieler,Martin,et al. "Predicting athletic success: Factors contributing to the success of NCAA Division I AA collegiate football players." Athletic Insight 9.2 (2007): 22-33.

[5] Brook,Stacey L. "An estimation of NCAA Football Bowl Subdivision demand as a two-part tariff." Managerial and Decision Economics 40.1 (2019): 79-83.

[6] Broyles, Kevin E. "NCAA Regulation of Intercollegiate Athletics: Time for a New Game Plan." Ala. L. Rev. 46 (1994): 487.

[7] Parkinson, Jerry. Infractions: Rule Violations, Unethical Conduct, and Enforcement in the NCAA. U of Nebraska Press, 2019.

[8] Kvam, Paul H., and Joel Sokol. "Teaching statistics with sports examples." INFORMS Transactions on Education 5.1 (2004): 75-87.