Spring 2024

Gridiron Insights: Predicting Gameday Outcomes Through Regression Analysis in College Football

Ethan Reyes

Justin Beal

Shawnee State University

Gridiron Insights: Predicting Gameday Outcomes Through Regression Analysis in College Football

By: Ethan Reyes & Justin Beal

A Thesis Submitted to the Faculty of Graduate Studies
In Partial Fulfillment of the Requirements
for the Degree of Masters of Science: Mathematics

July 28, 2024

Accepted by the Graduate Department

7/29/2024

Graduate Director Date

The thesis entitled 'Gridiron Insights: Predicting Gameday Outcomes Through Regression Analysis in College Football' presented by Ethan Reyes & Justin Beal, candidates for the degree of Master of Science in Mathematics, has been approved and is worthy of acceptance.

7/29/2024	
Date	Graduate Director
7/28/24	
Date	Student 1
28 July 2024	
Date	Student 2

Abstract

This research investigates the predictive power of the changes in spread, over/under betting lines, and home field advantage in determining whether the favored team in a college football betting market will cover the spread. The study examines three key factors: the change in the betting spread, the change in the over/under line, and the home-field advantage of the favored team. Using a comprehensive dataset of betting data from Draft Kings and Bovada, the study uses logistical regression techniques to analyze the relationship between these variables and the favored team's performance against the spread. Our findings indicate that fluctuations in the betting spread and over/under lines, combined with the home field status of the favored team, do not provide a statistical significant predictive insights. The results demonstrate that these lines can are not effective when utilized to predict the likelihood of the favored team covering the spread, highlighting efficiencies in the betting market. This research contributes to the understanding of sports betting dynamics and offers practical implications for bettors seeking to improve their wagering strategies through data-driven approaches.

Acknowledgements

I would like to acknowledge several individuals whose support and guidance were instrumental in the success of my research. Firstly, I extend my deepest gratitude to Professor/Advisor Dr. Darbro for his unparalleled wisdom and mentorship throughout the research component. I am also thankful to my partner in the study, Ethan Reyes, for his collaboration and dedication. My family provided unwavering support and encouragement, for which I am immensely grateful. Lastly, I appreciate my colleagues at work for their understanding and assistance during this journey.

– Justin Beal

I would like to extend my extreme thanks those who have helped along this journey and who have made many sacrifices to help me get to where I am today. Without the support of family, friends, and faculty this would not have been possible. To Dr. Darbro, thank you for your continued support and mentorship. Thank you for pushing for excellence and the betterment of my understanding in statistics. Thank you to my family who have been incredibly supportive of my dreams and have made sacrifices to help me along the way. Thank you to my partner Justin Beal for his commitment to excellency. Finally, thank you to my friends and colleagues who have encouraged me to pursue my masters and their support along the way.

– Ethan Reyes

Contents

List of Tables

List of Figures

1 Introduction

Chapter 1 will adequately equip the reader with a basic understanding of the intricate pieces that come together for the full study. For example, it will provide a concise overview of the literature review, delving into key articles pertinent to this study. The primary objective of this study is to determine if using variables such as power ranking, opening and closing money lines, the opening and closing spread and predicted winner to determine the result of a National Collegiate Athletics Association (NCAA) football game and ultimately whether or not the favored covers the spread. Before diving into the complexities of the study, a broad and general outline will be given. For instance, the significance of the study will be addressed along with the purpose and primary research questions involved. To give the reader a better understanding of how the study was conducted it will also discuss the research design. To ensure the reader is prepared for the remainder of the study a definition of terms will be provided. Finally, this chapter will also provide the theoretical framework used to build this study as well as address any limitations, assumptions, and scope of the study.

As indicated in the title, the study consists of using specific variables to predict the outcome of a college football game and if the favored team covers the spread. College football typically sees an estimated $8 billion gambled per season which turns out to roughly two thirds of a billion dollars being gambled each week for a twelve week season (Dodd, 2022). Even a small payout of 0.01% would yield a lucrative amount of $100,000 so it seems natural that one would want to discover how to predict outcomes in hopes of getting such a payout. However, there are a plethora of different ways to gamble, such as a parlay, betting on individual player stat lines, as well as betting on scores throughout the game. The type of gambling that will be researched in this study will be money lines and spreads. Furthermore, the focus will be to use the two types of aforementioned betting lines, along with a few other variables, to predict the outcome of a college football game.

1.1 Background

A brief overview of the relevant literature review is provided over the next pages but for a full in depth analysis, see chapter two of the study. The novel approach used in the study is unique in that it involves using betting lines to predict the outcome of a football game. Although there is minimal literature regarding using spreads and money lines to predict the outcome of a college football game, there is evidence to suggest it is a reasonable approach for gambling. Alternatively, there are other methods that have been studied for its predicting power specifically, reverse line movement betting. The process involves tracking the movement of the money and betting when the money lines movement in favor of the underdog reaches a certain threshold. The literature does not support reverse line movement betting being profitable or accurate. While it may not directly deal with predicting the outcome of a game, there is some literature giving possible explanations for bettors over/underrating the major conferences during inter-conference play. One of the final pieces of literature presented will be an article predicting the outcome of a soccer game using artificial intelligence. Even though this piece of literature is for a different sport entirely, there are still some takeaways regarding the methods used and the capabilities of artificial intelligence that are applicable to the study.

The first article subject to examination is one by (Cox, Schwartz, Van Ness, & Van Ness, 2021) which analyzes the betting line spread to predict the outcome of regular season games versus bowl games. The research conducted in the aforementioned article shares similarities with the study being conducted in this thesis but also has some major differences, both of which will be discussed. The data in (Cox et al., 2021) breaks the games into a few different categories, two of which are FW3 and UL3 which keeps track of a three game win or loss streak respectively. The interesting part of this variable is that there are often times when numerous different schools fall under neither category. However, it is arguably one of the stronger variables used for predicting an outcome. After all it is not often someone would bet on 10-0 Alabama team to lose to University of Tennessee Chattanooga nor would someone bet on 2-8 Louisiana Monroe Warhawks to win against #9 Ole-Miss. One important statistic the article uses is the “average number of bets placed for each game as well as betting imbalance". Including this in the model helps them understand the betting market of the game they are predicting. Overall, the results from the article by (Cox et al., 2021) support the concept of using money lines to predict the outcome of a college football game.

Since this study incorporates predicting if a favored team will cover the spread, it seems necessary to analyze the spread and what possible factors can cause changes in the spread. In particular, an article by (Coleman, 2017) seeks to analyze the validity of betting lines and discusses how travel time can cause changes in the spread. An interesting result the article found was “teams in arid regions win against the spread a significantly greater portion of the time when the visitor is from a humid region" (Coleman, 2017). It is clear from the study that weather indeed plays an important role on the result of the game. While the result of the game is undoubtedly important, the more applicable portion of this finding is the result that “the visiting team travels one time zone to the east, the home team scores a statistically significant 1.20 points more than expected" (Coleman, 2017). Although 1.2 points may not seem that much, it can certainly be the difference in making millions or losing everything. For example, consider if someone bet on the home team A to cover the spread of -8, then knowing the previously mentioned result could significantly impact decisions on placing bets as people would feel more confident that team A will cover the spread if team B is visiting from across the country. In summary, the results presented in this article offer valuable insight to line movement and reaffirm line movement to be used as a predictor in this study.

Spread betting is not unique to the college football betting market and in fact is used widely in other sports such as basketball, soccer, and baseball. An article by (Humphreys, 2010) offers insight for explaining movements of the point spread. Since the betting market relies on consumers to place bets on both winning teams and losing teams, they will sometimes move spreads to make it more attractive to bet on the underdog or the favorite team. Specifically “shading point spreads depend critically on the relative number of informed and uninformed bettors in the market" (Humphreys, 2010). In this context spreads will be moved so that uninformed betters will make bets on the “wrong" choice thus giving the betting agencies a way to cover paying out the potential winnings. While the findings of this article are based on predicting the outcome of an NBA game, there is still an application to the work presented in this study, namely, spread movement. Understanding spread movement is imperative to understanding a critical aspect of this study.

While this study will use money lines to predict if a favored team will cover the spread, it is not a novel approach to use money lines to predict the outcome of a game. For example, the article by (Berkowitz, Depken, & Gandar, 2018) analyzes different procedures for establishing win percentages from money lines. With a total of seven methods analyzed, the results indicated the two highest did not different significantly from one another. Furthermore, the first method involves “defining bets and returns for favorites and underdogs" as two separate equations used in the model to predict the outcome (Berkowitz et al., 2018). The second model also uses two equations for both the favorite and underdog but differ slightly in how they are defined. Since those equations and their derivations are not the main focus of this study, the full models and equations will be left for the reader to explore. It is worth noting that this article does not use the models to predict college football but uses it to predict the outcome of any particular sporting event. Arguably, since the article has applications for sporting events in general it proves the power of using money lines as a predictor. Moreover, the results of this study clearly justify the use of money lines as a predictor for a favored team to cover the spread.

Thus far the articles presented have a direct correlation to study presented. However, there is one such article that is not directly linked to this study but could potentially have impacts in the future depending on the field progresses. The article by (Guan & Wang, 2022) uses a neural network model and machine learning to predict the outcome of soccer games. While framework and concepts used and discussed in the article are applied to soccer matches they could potentially be applied to predicting a college football game. Naturally, this is easier said than done, but the potential advances for the field by applying this study are immense. The sample data in the study that was ran through the algorithm took inputs such as team goals, passes, offside, steals, possession rate and a few other team statistics. The data is then processed using a machine learning neural network to obtain a predicted winning rate. Obviously the team statistics recorded are taken from soccer but they could just as easily be college football statistics. Without overindulging in machine learning, one significant change that would need to occur to make this article applicable to this study would be the weights used in the machine learning process. These would have to be recalculated to emphasize the more important team statistics of a college football game like points allowed and win/loss record but give less emphasis on stats such as time of possession. As a whole, even though the article chose to analyze soccer, the main ideas will hopefully be implemented and designed to predict the outcome of a college football game.

1.2 Problem Statement

In the betting world, with an emphasis on NCAA football, predicting the outcome has been at the forefront of the gambling industry. To predict a game would require a hefty amount of statistical analysis and the size of the data in consideration would almost be insurmountable. In comparison, determining if the favored team will cover the point spread effectively conveys the same core idea. Therefore it remains to be analyzed whether it is possible to predict if the favored team will cover the spread. The study presented in this thesis specifically uses the opening spread between two teams, with the baseline of the favored team to win, in order to predict whether or not the favored team will cover the spread.

1.3 Purpose of the Study

This study will focus on the movement of the opening spread up until game time with analysis showing whether said movement is a significant predictor of the favored teaming covering the spread. Predicting any outcome of a contest, whether it be cards or associated with sports, was a main reason, if not *the* reason, for the subject of probability and its connection to statistics. The purpose of this study will be to examine and analyze the movement of the opening spread up until game to time to predict whether or not the favored team covers said spread. Additionally, this study may shed more light on the popular trend of the "transfer portal" and its significance with calculating the opening spreads and favored teams.

1.4 Research Design

There are a lot of moving parts incorporated into this study such as the money lines, point spreads, final score, home field team, power ranking and win probability. Luckily, there is one website that contains a vast majority of the data and it is conveniently titled: (CollegeFootballData.com, 2023). With historical game data dating back to 1869, this seemed to be the most ideal site to use to gather information. Even though the game results go back over a century, the archives for money lines and spreads only date as early as 2021. The youthfulness of the historical data is not of concern since legal gambling of sports was only recently allowed to resume back in 2018.

The historical money line and spread data found in this site is cited based on where the information comes from. For example, there are up to four sources for each game listed in the database. The two primary sources for the historical money lines and spreads are cited to be taken from DraftKings and Bovada. To understand why each site has different money lines, consider the following explanation. The respective sites wish to remain competitive yet profitable at the same time so naturally each entity will adjust their odds to ideally attract customers. However, any further analysis is out of the scope of this study but it was necessary to at least be mentioned to the reader.

Fortunately, the method to gather the necessary data used in this study is not complex but just tedious. The historical data is archived at (CollegeFootballData.com, 2023) and can be exported to a .csv file for ease of analysis. Favored team, opening money lines and spreads as well as closing money lines and spreads, score of home team, score of away team, can all be exported in one .csv file. The power rankings of teams will have to be exported in separate .csv but will need to be matched with its respective game/team found in the first file. The remaining variable to be gathered is if the favorite team covered the spread which can be calculated using a simple excel formula. Since the first .csv file contains the scores of the games it will be easy to write a formula that can be dragged down and applied to all games in the data set. If the favorite team covered the spread, a 1 will display in the column, if the favorite team did not, then a 0 will remain in the column. The variable is recorded in a dichotomous fashion for ease of analysis.

It seems necessary to disclose that the data collected is from a 3rd party not associated with DraftKings or Bovada. Even though the data is vetted by an administrator it is still open source. Money lines and spreads and their fluctuations can actually be viewed live on the website until their closings. The website cites its sources to clarify where the numbers are originating from but that does not ensure the data is completely whole and accurate. Small instances of human error, such as transposing two numbers or similar types of errors, are likely to occur and should be considered when discussing the origins of the data taken and used in the sample for this study.

The optimal statistical approach for analyzing the data set is linear regression. Linear regression was found to align the most ideally with the goals of this study. The linearity assumption of linear regression fits perfectly into the study since the expected result is that the stated variables will predict if a favored team will cover the spread. Furthermore, the model will also assess the strength of predictability for each variable. The model also allows for testing of different assumptions made about the study. As discussed in a later section, the assumption of independence of observations and errors can be tested effortlessly. In conjunction with other aspects of the study it seems most appropriate to use linear regression since there are both categorical and numerical variables being considered. Finally, the ease of interpret ability and simplistic implementation in the statistical software R, which is what was used in the study, is another driving factor behind using logistic regression.

1.5 Ethical Considerations

While college sports have been around for decades, gambling on them was only made legal as of 2018 with the Professional and Amateur Sports Protection Act (PASPA) being removed (Bengel, McCarriston, & Staff, 2023). The immense amount of wealth being gambled per season (nearly $8 billion as mentioned earlier in the introduction) brings rise to a few ethical considerations to be discussed. A natural question of sports gambling is if the game is rigged for certain a outcome or not. The response to this has a two-fold motivation behind the reason for rigging the outcome of a game. First, rigging of a game could be done for the sake of creating more attendance of future games for the winning team. For example, Alabama having a winning season is more profitable because people want to pay more to see their team win. Whereas if Alabama had a losing record the attendance would likely fall for games later in the season. The second aspect is the potential payout of the outcome. It is possible someone with a lot of money to win could offer to share the payout with players who throw the game or with referees who rule in favor of the team with a payout. Thus maintaining sportsmanship is imperative to keep the integrity of the predictions.

Another ethical consideration is the payout of the players. As discussed earlier it would be dangerous to give a portion of winnings to players in fear of them throwing games. However, it should be considered the money being generated would not exist without them or their athletic performance. Thankfully, nowadays NIL deals potentially solve this problem by allowing athletes to be paid for name, image, and likeness, hence the acronym NIL. Without NIL deals existing it would be a very slippery slope to allow people to amass fortunes based on the hard work and performance of many college athletes who would not see a single penny from the payout.

It is necessary to disclose to the reader that this study does not encourage gambling or justify any risks taken associated with gambling. The findings of this study may or may not find a strong method to predict the outcome of a college football game. However, it is left to the discretion of the reader to use these findings for gambling purposes. The gambling discussed and tested in this study are for academic purposes only. The authors of this study do not stand to profit financially in any way based on the results. Furthermore, the reader agrees to release the study and any persons directly associated with the production of the study from any financial liability stemming from their own actions. Please gamble responsibly.

1.6 Theoretical Framework

A better understanding of human behavior is necessary for piecing together the complex intricacies of gambling and its motivating factors. The academic field of behavioral economics offers insight on explaining the intricacies and motivating factors. It "began as a purely academic attempt at modeling irrational consumer choices, thereby challenging the notion of the rational consumer of traditional economics" (Reed, Niileksela, & Kaplan, 2013) and eventually it grew into the discipline that it is today. With applications in other academic fields such as psychology and finance it proves the versatility of the field and therefore justifies its use as the theoretical framework of this study. In the context of this study, behavioral economics will be used to offer an explanation for the movement of betting lines. For example, there are people who will bet on their Alma mater no matter how good or bad the team is because of their loyalty to school or to the program. People might also choose to bet by only looking at the odds and betting on the underdog in hopes of receiving a lucrative payout. Two similar types of gamblers are the professional gambler and the amateur sports better. An amateur sports better might be an individual looking at team statistics such as yardages, type of offense, and average star of recruits among other things to base their decision off of. In comparison professional gambler uses heavy statistical analysis and modeling to run simulations that will predict the outcome of a game. Arguably, the most interesting consideration are the professional gamblers who run statistical simulations on the games to predict the outcome and bet accordingly. These professional gamblers will put upwards of a million dollars on certain teams based on statistical analysis. Behavioral economics offers a much more in depth explanation for such types of betting than just individuals being driven for monetary success.

One article conducted a survey to determine the driving motivations behind individuals and their desire to gamble. Creatively titled *How rational is gambling?*, the article seeks to understand gamblers and continued desire to gamble “despite the fact that they lose money on average" (Stetzka & Winter, 2023). Clearly financial gain is not the only motivating factor, otherwise people would not be betting if they are making profits from their endeavors. However, the article also explains that the behavior of gamblers who are indeed making a profit closely resembles that of financial investors (Stetzka & Winter, 2023). Furthermore, there is evidence to suggest there is a rational motivation for certain gamblers and the types of betting they indulge in. Specifically, the gambler whose betting options emulate that of a diverse financial investment portfolio are the ones who have rational motivations for gambling. The results of this article clearly articulate that there is a distinct separation between gamblers who are making a profit and those who are not. The authors come to the conclusion that the ones who are profitable are using a pragmatic approach to their betting decisions. In contrast, the results indicate that the ones who are least profitable and “lose control" and therefore tend to make irrational decisions with their betting choices (Stetzka & Winter, 2023). Overall the article offers valuable insight into the mind of a gambler which is imperative to understanding the movement of betting lines in this study.

While defining possible explanations for the motivations of gambling it is also helpful to consider the demographic of gamblers. An article by (Snaychuk et al., 2023) further expands on the idea of behavioral economics by diving into the demographics of sports bettors. Analysis of the data gathered in this article concludes that the majority of sports bettors are male. Furthermore, it also found the following attributes to be most common: employed, young in age, single relationship status, and are highly educated (Snaychuk et al., 2023). The goal of the article is to identify significant risk factors of those who engage in gambling. Even though not everyone who engages in sports betting is considered to be addicted to the point of risking entire life savings, it is still important to know the risk factors to prevent loved ones from falling victim to addiction. Another interesting result mentioned in the study is that “sports betting is generally perceived as low risk by those engaged in it" (Snaychuk et al., 2023). In most cases this can certainly be true. For example, a casual sports bettor who designates for themselves an allowance as a form of entertainment would be considered low risk. Similarly, professional gamblers would also consider their activities low risk based on the simulations they use to make their betting decisions. However, there are certainly instances of those who bet their entire life savings and naturally this is not considered low risk.

1.7 Assumptions, Limitations, & Scope

Independence of errors is an assumption made in this study. Fortunately, the residual plots indicate that dependence of errors is not a concern and therefore ensures the validity of the logistic regression model being used. Similarly, independence is assumed for each of the observable instances in the collected data set. Namely, any game in any given week is considered independent from all other games. Intuitively this makes sense, the result of a week one game with Ohio State vs. Indiana, has zero effect on the result of a week eight game with San Diego State vs. Nevada. It is also worth mentioning that statistically, the result of week one Ohio State vs. Indiana is also independent of the result of week eight Ohio State vs. Wisconsin. If this study considered variables such as injuries or perhaps covered multiple seasons, then one could argue there is a concern for violation of independence. However, since neither are applicable to this study, there is no concern for violation of the Independence assumption. The independence assumption for the observable points in the data set allows for the use of logistic regression which is critical to the study.

One last assumption made is the integrity of the sport and of gambling. As mentioned in 1.5 Ethical Considerations, it is not out of the question that there might be instances of tampering in certain games based on the potential payout. One such incident occurred recently when a quarterback from Iowa State was accused of gambling on Iowa State football games (APNews, 2023). Although there is no concrete evidence to suggest this particular quarterback threw a game to receive a portion of a payout, it is still worth mentioning to show college football is not immune to such scandals. However, for the sake of simplicity, integrity of gambling is assumed to remain unbroken. Furthermore, sportsmanship or the integrity of the sport is also assumed. As seen in the 2023 season the integrity of the sport is not always maintained. Namely with head coach Jim Harbaugh being suspended for sign stealing from opponents (Meyer, 2024). It is possible there are other similar situations actively happening but not yet caught and revealed to the public. In the interest of simplicity such actions are not considered in this study and the integrity of the sport is assumed to be held true.

Limitation of data availability is a minor concern of this study. As discussed earlier in the chapter the two main sources from which the records of money lines and spreads, originate from Bovada and DraftKings. There are clearly more than just two sites used for betting on money lines and spreads. However, the website used to gather the data for this study only has the two aforementioned sites recorded in its archives. With an overwhelming number of sites to gather betting lines it would be possible to record the same data used in this study as the season progresses. For example, in the upcoming 2024 season, one could visit six different sites and record their opening and closing money lines. The limitation of data availability would only be concerning if there was only a website then historical data recorded from. Therefore having two websites, Bovada and DraftKings, does not hinder the results of this study. Moreover, this study does not consider the fluctuations in the opening and closing spread but rather movement is defined to be the difference between the initial opening and final closing spread.

Another limitation of this study are unmeasured factors which can be both a recorded stat line or something intangible such as team chemistry. It would be too tedious to list all of the stat lines not considered or used in this study but here are few examples. Reporting the number of injuries per team could possibly have some correlation but to truly find out, would divert from the main focus of the study. One issue with reporting the number of injuries a team has in any given week is having to note how many are starting/key players. The weather of a game is another aspect that could be measured and analyzed since teams from different regions of the country are more acclimated to the weather they train in. Another intangible measure would be styles of coaching. It would be almost impossible to record this as some form of data but undoubtedly has some impact on the result of the game. Although these limitations are unmeasured factors, one could argue that they are actually embedded in the money lines and point spreads. For example, it is possible that people are either consciously or subconsciously considering these things when betting on the game.

The scope of this study is limited to the 2023 college football regular season. Furthermore the games analyzed are all division one teams. It did not seem applicable to attempt to predict bowl games since there would be additional factors to consider. For example, teams with players seeking to enter the NFL draft often sit out of bowl games to prevent the risk of injuries giving those teams a slight disadvantage in the bowl game. Additionally, to keep the sample size large enough for bowl games the study might have to include multiple years at a time. Furthermore, the variable of home or away would have to be disregarded for bowl games since they are played on at neutral location. Therefore to keep the scope of the study simplistic, bowl games are not included in the sample.

1.8 Definition of Terms

Money lines: Used to measure the likelihood of a particular outcome can be either positive or negative. A team with a -225 money means in order to make a profit of $100, an individual must place a $225 bet on that team. Conversely, a team with a +225, means that a bet of $100 will result in a $225 profit. Additionally, the team with -255 is anticipated to win the match up.

Point Spread: Refers the margin by which a team will win. A spread of -8 for team A means that team A is anticipated to win by a margin of 8 points.

Power Ranking: Teams can be assigned a rank of 1 through 25 each week of the season. The rankings are determined by the Associated Press Poll, who take into consideration the results of the previous week, strength of schedule, and other various factor to determine theses rankings. Each week new rankings are released.

Favorite team: The favorite or favored team refers to the team that is anticipated to win against their opponent. The team considered the favorite is identified by a negative point spread or money line.

Covered Spread: To cover a spread a team must win by more than the closing spread. For example if a team wins by a margin of 8 points and the closing line was -8, then the team has covered the spread. Conversely, if the favorite team wins by 7 points on a closing spread of -8, then the team did not cover the spread. Finally, if the favored team loses by any margin on a closing spread of a negative number, then the team did not cover the spread.

Reverse-Line Betting: When a betting line, the spread in this case, moves in the opposite direction of where the popular bets are being placed. Usually number bets can be tracked in terms of a percentage. For example, 80% and 20% for a particular spread, so the popular bets are with the 80%. Reverse line would then move in the direction of the 20% bet rather than the 80%.

1.9 Summary

In summary, chapter 1 serves as a brief overview of the study presented in this thesis. It will also set the stage for a more in depth look at sports betting and its many models using spread and/or money-lines to predict certain outcomes of NCAA football games. This particular study will look at the movements of the spread between their initial opening and game time. In addition to this movement, the home money line and visitor money lines will also be included in the analysis of whether the favored team covers the spread.

Data analysis will involve logistic regression techniques. As mentioned above, independence for week to week games will not be an issue and the independent variables will be the movement of the opening spread up until game time along with home/visitor money-lines. These will be the basis of the study to then predict the dependent variable of the favored team covering the spread. The study will be using the Theory of Behavioral Economics to look at how the spread moves between opening and game time. As it will shed some light on how “big bettors" look to move the spread line in order to place a bet that their particular model predicted to be the best line for predicting their outcome. This information could be used to further the understanding of this spread movement and its statistical significance in predicting the outcomes of a wide range of sports in future studies and models.

2 Literature Review

Chapter two's aim is to present a comprehensive literature review of the extensive research conducted on sports gambling. This process entailed a detailed analysis of each article and its references. This was accomplished by employing a meticulous search strategy, exploring various databases and utilizing diverse search terms. The focus was to identifying pertinent theories, research outcomes, significant models, and data to construct a more nuanced and well-informed understanding of the sports gambling, specifically NCAA collegiate football during the 2023 regular season.

2.1 Modern Sports Betting and Bookkeeping

Today's gambling world, and more specially, that of American football is made up of three types of betting "lines": the point spread, the over/under, and the money-line. These three "lines" will be discussed more in-depth in later sections, so for now "lines" will reference a simple bet being placed using one of the three types mentioned above. Modern bookkeeping and their keepers have found a way to create a market where they are removed from any significant loss. As the bookkeeper they establish a "line" where their goal is to have 50% of the bets made on both sides of said "line". This way those that win are paid by those that lose. The bookkeeper makes a commission (aka juice/vigorish) on the losing bets, generally 10%. That is to say, if a $100 bet is made and loses, the bettor must pay an additional $10. Whereas if the bet is won, they will receive $100 on top of the initial bet. So, one can see if a bookkeeper takes 2 bets, one for each side of the line, the $100 from the loser will go to the winner and the commission of $10 will go to the bookkeeper. For our current research it is important to note that the "lines" offered by this research are virtually identical to those at any bookmaker online or at Las Vegas casinos. In practice, individual bookmakers are not actively setting prices but following the lines set by Las Vegas casinos and their associated odd makers (Levitt, 2004).

These "lines" are sometimes created using a statistical model based on what is known as the Efficient Market Hypothesis (EMH). In general, EMH states that there are 1) bettors and those who take bets / bookkeepers, 2) both are assumed to be rational "profit-maximizers", and 3) who both have free access to virtually all pertinent information about the "line" in question (Pankoff 1968). Under this model, bookkeepers can profit without much risk by creating a "line" that attracts equal bets on both sides (Paul & Weinbach, 2008). But as modern research has consistently shown, these lines must be adjusted as the number of bets on each side will almost always be skewed one way or the other.

With regards to betting on collegiate American football, Sinkey and Logan have shown that betting markets and their bookkeepers are less efficient. This is perhaps due in large part to the nature of it being defined as an "amateur" sport and so the rules are not as strict with respect to a school's obligation to share news about players as is the rule in the National Football League (NFL). This is not so when it comes to the coverage that collegiate teams receive throughout the year. As with all NFL teams there is significant media coverage but this only so for a very few collegiate teams (Sinkey & Logan, 2014) Their study showed that the lack of media attention only magnifies this inefficiency. Although bookkeepers act as though the market is "efficient", as defined above, this strategy can be disastrous. Compared to the thirty-two NFL teams, collegiate football has over 830 teams/schools, this includes all divisions (Manuel, 2022). The study presented today will only include the 129 teams/schools in the Division 1 Football Bowl Subdivision (FBS). An almost three-fold difference between FBS and the NFL. As making it difficult for bookkeepers to create an efficient market for the 64 collegiate games per week versus only 16 in an NFL week.

2.2 An Overview on Sport Gambling Lines and Odds

Let us now present this study's definition of betting "lines" and the significance of each. There are a few types of betting features on many of the modern gambling apps for NCAA football. The ones discussed in this paper will be the spread, the over/under, and the money. An example for each will be sufficient to show how each will be presented. The "spread" line is one that involves the picking the winner of a game against the "spread" of the winner and losing teams scores. For instance, suppose that a bookkeeper or betting website sets a betting line, in this case a spread on the Michigan versus Ohio State game where Ohio State is favored to win by ten points. Suppose that the game occurs and Ohio State wins by 31 points (Sinkey & Logan, 2014). The "over/under" line is a bet placed on the total sum of the final scores between two teams. A bettor decides is this total sum will be either "over" or "under" the bookkeepers line. So if the previous examples game between Michigan and Ohio State has a score of 14-45, and a bookkeeper registers a line of 63.5 points any bettor that chooses the "under" will win as the total points for the game is 59 points and therefore under the bettors line and those who chose the over will lose.

Lastly is the money line bet. Betting on the money line is one of the simplest bets and it's popularity has only grown over the years. A money line bet is simply betting on who will win the contest, regardless of spread or over/under lines (Schaefer, 2011). As mentioned, the money line does not consider the score for each team but bettors must look at tow numbers of positive and negative numbers. The negative number indicates the favorite. For example, if a team has a money line of -150, you would need to bet $150 to win $100 (plus your initial stake back). The negative number tells you how much you need to bet to win $100 profit. The positive number then indicates the underdog. Using the Ohio State-Michigan example, if a Michigan has a money line of +150, a $100 bet would yield $150 in profit. For the Ohio State money line of -150, a bettor would need to wager $150 in order to win $100. The positive number tells you how much you would win on a $100 bet. In most cases, money line bets don't have a provision for a tie or push. The bet is resolved based on the outcome of the game, either one team wins or the other. Money line bets are popular among casual bettors due to their simplicity, and they're often used when betting on favorites or underdogs with high odds. However, keep in mind that while they're simple, they can still carry risks, particularly when betting on heavy favorites or significant underdogs.

2.3 Variables Influencing the Outcome of a Game

Across all sports there are a plethora of different variables that can influence the outcome of a game. While not every model used to predict the result of a match includes these variables as a predictor, they inevitably have some affect on the result. For example, one clear such example is home field advantage. An article by (Krieger & Davis, 2023) investigates this idea of home field advantage. The authors suggest that it is a simple as home versus away but rather there are numerous intricacies that factor into it. To start with, they note that majority of research has indeed confirmed the presence of home field advantage but their findings seek explain that the it is not "a simple binary factor (home or away)" (Krieger & Davis, 2023). The article specifically mentions factors such as location, weather, fan participation, and intimidation of location history among others in its explanation for home field advantage. The article uses pre-COVID-19 and post-COVID-19 attendance numbers and scores to examine the impact of fan participation. Ultimately, it concludes that fan participation has a significant impact on home field advantage. Moreover, closing lines favored the home team by only "one point (.99) as opposed to the 2017-2019 average of 2.32 points" (Krieger & Davis, 2023). Based on this significant shift in closing liens it appears that bookmakers also understood that with lack of attendance, home field advantage would not be as strong as it would be with fans. It is important to note that the sample for this article is professional American football, National Football League (NFL), and not college football. However, the results and findings are still applicable to this study. The difference between the study being conducted and the article presented is home or away is treated as a binary factor for this study. However, the physical and intangible influences previously mentioned of a home field advantage are embedded into a gamblers decision when considering who to place a bet on.

Not only does the location of a game matter for home and away purposes but there are other aspects that play a role as well. Things such as the climate of the location, the weather that specific day, and the distance travelled by the away team all significantly impact the outcome of a game. There is evidence to suggest that teams from an arid climate are more likely to win when hosting a team from a humid climate region (Kuester & Sanders, 2011). Consequently it also draws the conclusion that it is indeed a profitable strategy to bet on an arid region team when they are hosting a humid region team. The article also examines the other three possibilities of arid hosting arid, humid hosting arid, and humid hosting humid. The two least interesting combinations of arid hosting arid and humid hosting humid win against the spread 50.2% and 50.7% of the time respectively. Even though humid hosting arid regions win against the spread 48.9% it is not significant enough to draw any sort of conclusions about betting or about an inherent advantage to a specific regions. Finally, arid regions hosting humid regions won against the spread in 56.6% of the sampled games (Kuester & Sanders, 2011). While the 56.6% is strong enough to conclude that it is profitable to bet on arid hosting humid regions, it is not statistically powerful enough the conclude that their is an inherent advantage due to the climate. A possible explanation might be the time of the year the games were played since each region tends to have more extreme weather at the beginning and end of the season. While adjusting to the change in climate when traveling undoubtedly has an effect on the outcome of a game, it remains to be proven if there is an advantage for one region or the other.

One article found a correlation between the number of time zones traveled to a higher chance of losing further expanding on the idea of traveling teams being at a disadvantage. Specifically, "when the visitor crossed one or more time zones to the east" the home team scored more points than originally predicted (Coleman, 2017). The author suggests it is the loss of time when travelling eastward that causes the visitor to be at a disadvantage. Finding if there is a particular range of distances that has the most impact on a game could be a potential future study branching from the article. For example, knowing if teams that travel eight to ten hours prior to game day are at a disadvantage would have significant impact on the college football industry. In addition to the number of time zones travelled the author also suggests that later in the season the more sensitive the game is the time zones travelled phenomenon. Specifically, road underdogs are less likely to win when traveling at least one time zone to the east (Coleman, 2017). There are plenty of possible explanations for a late season underdog having a lesser chance of winning, one of which being fatigue. Both mental and physical fatigue more than likely each contribute to the explanation. Regardless of the explanation, the results of this article justify the usage of having home or away as a variable in this study.

More specific to college football, the conferences college's are from has a bearing on the outcome of a game. Presented next is an article by two authors that discuss how a college's conference is undervalued when betting (Moore & Francisco, 2019). Quite similar to a few of the articles already, this article mentions that the results found are not statistically significant enough to prove teams from a power five conference are more likely to win in an interconference game. However, the authors did suggest that betting on power five schools to cover the spread in general is profitable. For context, a power fiver conference refers to the following conferences of college football: Big 10, Big 12, Atlantic Coast Conference (ACC), Pacific 12 (Pac-12), and the Southeastern Conference (SEC). Teams that win in any of the respective conferences automatically get a bid to play in one of the major bowl games. To an avid college football fan it is obvious why betting on teams from these conferences to cover the spread are profitable. Teams from these conferences are typically powerhouse schools and a majority of the time the top 25 ranked teams consist of schools from these conferences. Therefore it is reasonable to believe that betting on these teams to cover the spread is profitable.

Another interesting factor to consider that might influence an outcome of a game is whether a team had a bye-week the week before or not. It is commonly thought that bye weeks can help a team get prepared and get some well needed rest to adequately perform for the remainder of the season. However, research suggests that the opposite is true and in fact teams are more likely to lose after having a bye week (Howington & Moates, 2017). The results also indicate that the odds of losing for teams coming off a bye week in the latter half of the season increases even more so. As shocking as this is, the author presents a few reasonable explanations for this being case. The first being having a bye week scheduled the week prior of facing a strong or higher ranked opponent. In this case it could be purely coincidental that the bye week team was going to lose anyway to the stronger opponent. Another possible explanation given is that the bye weeks break the flow of focus and therefore cause the team to not play their best game. Unfortunately, this article did not study the profitability of betting on teams who have just taken their bye week. As seen in multiple articles already presented, statistical significance does not always equal profitability. Therefore careful consideration should be taken if choosing to place a bet based on the bye week of a team.

Rivalries are always an attention grabber across all sports and in college football there is an entire week dedicated to rivals battling it out on the playing field. With certain match ups dating back to the inception of college sports, one may look to the past years game results to predict who will win. However, there is evidence to suggest that "rivalry games are harder to predict" than regular season games (Winfree, 2020). The model used by authors correctly predicts the winners of all games at 73.7% but when specifically predicting rivalry games the model only was able to predict 66.7% and full 7% lower. The author suggests these results imply rivalry teams are more evenly matched then other against other teams. Additionally, the results conclude the following: bowl eligibility and traveling on the road the prior week has no bearing on the outcome of game, playing a strong opponent the week prior helps home teams, and having a close game the week prior negatively impacts the team in the upcoming week (Winfree, 2020). The first two of the aforementioned results might considered to standard conventions since the idea of traveling a week before is under the perception of travel fatigued for the team. Similarly, not qualifying for a bowl game can give the perception that the season is already over and apathy starts to set in for the team thereby decreasing the desire to win. A possible explanation for a team to do better after playing a strong opponent might be the team was locked in which transitioned to the week after regardless of the outcome (Winfree, 2020). Finally, a close game having a negative impact on the outcome of a game might stem from two reasons. The first being a team's mental state being defeated and depressed after losing by a small margin. Conversely, the second reason could be teams who win get carried away with the euphoric feeling of winning a close game. Regardless, all of these results have some sort of impact on the outcome of the game and should therefore be considered when placing a bet on certain games.

While the factors found affecting home field advantage, or any topic covered so far in this subsection, in the previous articles are not recorded or calculated in this study, it is imperative to present them to the reader to understand they are taken into consideration when placing bets. Moreover, it could be argued that bookmakers have an inherent deep understanding of the factors influencing the outcome of a game. However, given how intricate and detailed these factors are, it is much more likely they have broad understanding of the concepts previously presented. It might prove interesting to study which of these factors has the most influence on the movement of betting lines but that would venture out of the scope of this study.

2.4 What Affects the Movement of Money-lines and Point-Spreads

So far the discussion of what moves the betting lines has taken place but not why betting lines move. A brief explanation would be that the entity hosting the betting lines, moves the lines to maximize profit and loss on each game. Working for these entities are the people directly in charge of deciding which way the line will move after seeing bets are called bookmakers. This concept is not exclusive to college football but all types of betting. Bookmakers can directly or indirectly influence the movement of money lines. For example, bookmakers can choose whether or not to accept certain bets and in particular they are cautious of accepting extremely large bets (Hodges, Lin, & Liu, 2013). Additionally, bookmakers can also initialize their betting lines to influence the flow of betting to minimize their own risk of payouts. Essentially, they would open their money lines to influence the public into placing bets in ways that make it profitable for the bookmaker regardless of the outcome of the event. In particular this could mean making underdogs have a lucrative payout and consequently making it less appealing to bet on the "obvious" winners. Although this article focuses on the horse race betting market, the general ideas discussed are still applicable to college football gambling. For college football this might take the form of bookmakers moving the point spread in one way or another to ensure that they will make a profit or at least break even regardless of who wins the game. Furthermore, since websites such as DraftKings and Bovada host multiple college games at once it is possible that they are willing to shift money lines for multiple games at a time. In other words, realizing a payout for game A would be a loss for them, they could shift the money lines of game B and C to ensure they would retain profit as a whole despite having one game resulting in a loss of revenue. In summary, bookmakers ultimately will make decisions to ensure their own profit which involves their decision to move money lines.

So far the literature presented has discussed in general bookmakers attempting to keep the lines profitable for themselves, but not much has been said as to how they do so. The next article presented will give insight on different strategies they use and present real numbers that they use to justify them moving the lines. For example, bookmakers tend to follow the 11 for 10 rule which means that an $11 winning bet only yields a profit of $10 (Vandenbruaene, Ceuster, & Annaert, 2022). Instead of receiving the full winning bet of $11 back the bookmaker will charge a fee for their services leaving the bettor with only a $10 profit. A bookmaker will also move the line to make sure at least 50% of bets are placed on either side. With 50% of bets on either side they guarantee they will cover the payout for winning bets from they money they win from losing bettors. To keep this even payout bookmakers might choose whether or not to take a position on the game to influence the flow of betting in their favor. This could mean shifting the line to influence more people to place bets on the underdog or vice versa. Consequently, the betting lines are a "forecast of the market's expectation of the game instead of the game outcome itself" (Vandenbruaene et al., 2022). In general the bookmakers could care less who actually wins the game their primary focus is ensuring they are maintaining profit as a whole across the games they are hosting. The findings of this article offer much insight on the inner working of bookmakers and explanations for line movement.

As discussed thus far, the money lines and spreads will shift one way or another based on the flow of bets coming in. One article seeks to explain how the flow of betting and large movements in betting lines can be influenced by the amount of media coverage of a game. More specifically, the authors chose to test "whether NFL games garnering less public attention are more likely to see larger movements in their betting lines" (Krieger & Davis, 2022). The results confirm their hypothesis that the less media coverage there is of a game the higher the chance of seeing large movements on the respective game's betting lines. Krieger and Davis suggest the lack of media coverage causes the large line movements in the following ways. One suggestion is that bookmakers tend to focus on games with most media coverage since those will have the biggest pool of money associated with therefore requiring more attention to detail to ensure profitability. Aside from just focusing on the bookmaker shifting to a more important game, they also suggest that the lack of information available because of the coverage can cause the bookmaker to improperly set the opening lines and therefore make adjustments while they are open to again ensure profitability. Lack of information to the viewers also plays a factor in these large movements in terms of the actual number of bettors gambling. For instance, if only a small number of bets are placed on the games with low coverage then it becomes more volatile to changes of all sizes and especially the big bets. While this article does test data from the NFL and no college football it still gives a generic overview for changes in betting lines.

2.5 Modern Professional Gambling and Other Statistical Models

Previously there have been attempts of home underdog gambling to maximize profitability of a bet on college football. A home underdog would be when the home team is not considered the favorite and is thus predicted to lose the outcome of that match. Using data ranging from 1976 to 2000 it was proven profitable to wager on a home underdog (Paul, Weinbach, & Weinbach, 2003). However, the results conclude that betting on home underdogs by a margin of 7 points but less than 28 points is not profitable. In fact, once over the 28 point threshold, it is statistically significant to conclude that underdogs are more likely to win in addition to being profitable. For underdogs between the margins of 7 and 28 point differentials, the results conclude this would be considered an unfair bet. An unfair bet refers to the idea that point spreads set by bookmakers do not accurately predict whether or not the underdog will beat the spread. It is precisely because of the unfairness of the bet that betting on underdogs in the aforementioned range is not profitable. Furthermore, the results also conclude that underdogs in general are considered to be efficient but "large underdogs (more

28 points) were found to violate a fair bet" (Paul et al., 2003). Betting on underdogs in general over the 28 point threshold is not profitable because the author suggests that enough informed betters are already doing so. In other words, the informed betters are keeping the betting margins at a level that suggests consistently betting at those levels would not be profitable for the reader. In summary, using an underdog betting model can be profitable depending on how it is done.

Furthermore, one article takes it one step further and suggests that betting on underdogs in general, not just a home underdog, is profitable and analyzes four different strategies for doing so. The four different strategies all involve betting on the underdog when the spread is greater than five points but each containing a different component to give it some sort of uniqueness worth testing. After testing each strategy, the authors come to a conclusion that it is profitable to bet on "underdogs who receive five and a half points or more from bookmakers and no fewer than two points from syndicates" (Badarinathi & Kochman, 1996). Essentially betting on teams in which the spread would show them to lose by a margin of five and half points or more would be considered profitable. It is still necessary to discuss these models regardless of how recent the models were developed or that they pertain to the NFL and not college football. It seems inconsequential the article is dated in 1996 considering how long both college and professional sports have been around. Moreover, the same model could just as easily analyze data from college football and not the NFL. That is not to say the results of which strategy is more profitable would be the same but rather proves the versatility of the model to be generalized to other data sets.

One article finds that because the potential to score zero points is critically underestimated, it is possible to exploit this result to profit off of college football betting lines. The underestimation of scoring zero points is called a market inefficiency. One article analyzes such an inefficiency and discusses the consequences. In particular the results find betting on games with significant underestimation can be profitable by betting on the over instead of under (Arscott, 2023). Censoring at zero, underestimating the potential to score zero points, could stem from biased ideas that a complete shutout victory is considerably harder just outright winning as it is almost seen as having a perfect game. A considerable amount of research has been dedicated to finding such inefficiencies and experts or well informed bettors will often be able to exploit them for maximum profit. One limitation of using this exploit is the necessary analysis required before concluding there is censoring at zero. Again, it is much more reasonable to simply know how accurate opening spreads and money lines are to determine if the favored team will cover the spread. By implying knowing if there is a high chance overall in teams covering the spread it reduces the necessary analysis to be done before placing a bet. In other words, this study will be applicable to all situations and generalized enough for even an amateur to better understand.

It is standard practice to examine or disclose which variables are considered biased or unbiased predictors when creating a model for game prediction and spread coverage. An unbiased predictor is considered any object variable used in the model. Specifically, things such as team statistics, win/loss records, player performance, and game conditions are considered unbiased variables. Conversely, things such as power ranking and historical performance would be considered biased predictors. One article decided to test the possibility of opening and closing lines being biased predictors. However, the results indicated that they are considered unbiased predictors (Gandar, Zuber, O'Brien, & Russo, 1988). While the data used to draw this conclusion was taken from the professional football betting market, the article still remains relevant through its analysis of spread betting. Trivially, the analysis of being a biased or unbiased predictor directly implies the usage of spreads being a predictor. Even though in the context of the article spreads are used to predict outcome of games, it still applies to determining if a favored will cover the spread. Extending spreads to predicting actual coverage of a spread is not that far of a leap considering that half of the job is predicting if the team will win or not. As a whole the results found in this article justify using opening/ spreads as a variable in this study.

Articles discussing profitability until this point have commonly discussed different game conditions or different money line conditions as indicators of profit. However, the next article presents a mathematical approach to maximize profit on point spread, money line, and over-under betting. For example, one theorem to maximize point spread betting profit states: "to maximize the expected profit of a wager, one should bet on the home team if and only if the spread is less than the $\left(\frac{1+\phi_h}{2+\phi_h+\phi_v}\right)$-quantile of m" (Dmochowski, 2023). The reader is encouraged to view this article and examine the theorems, corollaries, and their proofs for a better understanding, as the actual equations presented are not the focus of this study. The most important parts are the broad concepts and results presented in the article. Specifically, the concepts presented use statistics to determine what *will* be profitable to bet on using data from the present. Conversely, a majority of the articles previously discussed regarding profitability analyze historical data and state if one had bet in a certain way, it would have resulted in a profit. Understanding the difference between those two concepts is imperative to realizing how powerful such a model can be. Feasibility of using the equations and conducting the statistical analysis are the biggest limitation issues of the article. It should be noted individuals, such as professional gamblers, who have a stronger background in statistics and gambling would likely find the results more lucrative and suitable to their needs. As such, it is necessary and more feasible for an average person to analyze betting lines to predict if the favored team will cover the spread.

2.6 Summary

Overall the articles presented in this chapter justify the idea of using opening and closing spreads, home/away money lines, opening over/under and power ranking to determine if the favored team will cover the spread. The novelty of conducting such a statistical analysis is justified by the minimal literature surrounding the concept. However, as clearly stated previously, there is sufficient literature to justify the individual components of this study. For example, there is ample evidence to suggest that considering if the favored team is home or away is a necessary variable to include. While there are some inefficiencies in spread betting and money lines, there is evidence to suggest that exploiting them can be profitable for the gambler. In conclusion, this study has been adequately justified through its presentation of relevant academic literature.

3 Methodology

In this chapter, a detailed overview of the methods used to analyze the betting lines of the 2023 NCAA football season will be outlined and presented. The researchers aimed to determine whether these betting lines were statistically significant predictors of the favored team covering the spread. The research procedure, data processing and analysis, limitations, and ethical considerations are thoroughly outlined as well as the rationale for the research design. This chapter will also set the foundation for the results presented in chapter 4.

3.1 Review of Primary Questions and Hypothesis

Data collected from the 2023 NCAA college football season included the home/away opening and closing over/under lines, the home/away opening/closing spreads, the home/away team money lines, home/away end game scores, and whether or not the spread line changed. However, as the present study looks at predicting whether or not the favored team covers the closing spread, any data that is collected after the game will be contrary to the overall study. That is to say, this study will only be using the over/under lines, spread lines, money lines, and whether or not the spread changed up until game time.

There is a high demand in sports betting in knowing if a team will cover the spread or not. Being able to predict such an outcome leads to models being created to maximize profitability in betting. This study seeks to analyze specific pre-game variables to find if they are able to predict if a team will cover the spread. The main hypothesis posed is the following: H_o : In the 2023 college football seasons home money lines, away money lines, opening and closing spread, over/under, and the favored teams are significant predictors for FBS teams to cover the spread. H_α : In the 2023 college football seasons home money lines, away money lines, opening and closing spread, over/under, and the favored teams are NOT significant predictors for FBS teams to cover the spread. The secondary hypothesis to be tested is the following: H_o : In the 2023 college football seasons home money lines, away money lines, opening and closing spread, over/under, and the favored teams are significant predictors for FBS teams to cover the over/under. H_α : In the 2023 college football seasons home money lines, away money lines, opening and closing spread, over/under, and the favored teams are NOT significant predictors for FBS teams to cover the over/under. Naturally both hypotheses will use the same data set and consequently will draw conclusions based on logistical regression models to test them.

3.2 Setting and Participants

For the purposes of this study the sample is taken from the 2023 college football season. In particular the regular season games for teams in the Football Bowl Subdivision (FBS). However, since not every game played by an FBS team is always against another FBS team there are a few cases in which games involving teams from the Football Championship Subdivision (FCS) are considered. For example, FCS teams playing FCS teams were not taken into the dataset to be analyzed but an FBS team playing a FCS team was considered in the dataset to be analyzed. As such, games not involving an FBS Team are not considered for analysis. Furthermore, the money line, over under, and spread were taken from only two sites, Bovada and DraftKings, with no other betting site being considered for analysis.

A total of 1,085 games were collected as the total sample size for the study. As such a priori statistical power analysis using G*Power was conducted, yielding a result that 351 games was the minimum number of data points needed to conduct the experiment. With 1,085 data points collected, this requirement was far exceeded. As a result, the generalization of the findings will correspond directly to the college football betting market. However, certain threats to the generalization need to be addressed. The first threat to be considered is the data used only pertained to Division I FBS teams. Consequently, results may not directly apply to teams in the FCS or other divisions. Second, applying these results to bowl games would not be considered best practice since bowl games are commonly considered an entirely different category to handle (Winfree, 2020). The final threat to the generalization of the results is time. Since the data is limited to just the 2023 season, future drastic rule changes or coaching staff turnovers could and will affect the "dynasties" of college football teams. For example, the Nick Saban Dynasty at University of Alabama came to a close at the end of the 2023 season. Having multiple such instances occur in any given year in the future could change the application of these results.

3.3 Instrumentation

With a total of 2,935 entries in the database for the 2023 season, it became necessary to trim down any unnecessary data points to prevent unintentional false statistical significance and ensure genuine patterns were found. The following types of games were eliminated from the dataset. Firstly, games missing a money line, over under, or a opening / closing spread were deleted from the data set. Trivially, deleting those games from the dataset is obvious since they are the variables being used to see if the home team covered the spread or not. And secondly, games whose betting lines that did not originate from Bovada or DraftKings were also deleted. After using the previous criterion to sort through the unnecessary data it became apparent that the only two websites cited who had all the necessary information were Bovada and DraftKings. Although it would be interesting to see if any site besides the two had better insights in predicting spread coverage, it unfortunately cannot be determined with the data provided.

A common issue is lack of data but the contrary seems to the case for this study. Careful selection of variables is necessary to ensure quality statistical results stem from the data analysis. All of the variables, except one, were converted to be a dichotomous variable, that is, having a 0 or 1 in place to indicate a specific result. The first variable used, DichoSpreadCovered, describes whether or not the favored team covered the spread. This is the dependent variable that the model seeks to explain from the rest of the other independent variables listed and defined below. In terms of the statistical model, this is the Y_1 for our equation

$$Y_1 \sim X_1 + X_2 + X_3 + X_4 \tag{1}$$

The first independent variable, DichoFavoredHome (X_1), is if home team is also favored to win. Statistical significance with this variable would indicate that there is a correlation between the home team being favored and the favored team covering the spread. Thirdly, DichoChangeSpread (X_2), if there was a change in opening and closing spread lines. For example, if the spread at the beginning of the week decreases due to a star player's injury who plays for the favored team. Statistical significance with this variable would indicate that there is a correlation between a change in opening and closing spread. Fourthly, DichoChangeO.U (X_3), if there was a change in opening and closing over/under lines. Statistical significance with this variable would indicate that there is a correlation between a change in opening and closing over/under. Finally, the difference in home and away money lines was calculated by taking the home money line - away money line. The sign and size of the difference calculated are indicators of the margin of victory. For example a larger positive (negative) number would mean the away (home) team has a higher chance of winning the match. Testing the secondary hypothesis involves changing DichoSpreadCovered (Y_1) to DichoOUCovered (Y_2), while keeping the same four independent variables. College football betting lines are typically posted by early in the week before the games are played. However, the exact timing can vary depending on the prominence of the game. High-profile games or marquee match-ups, betting lines may be posted even earlier. Lines can also be updated throughout the week based on betting activity, injuries, weather, and other factors that might influence the outcome of the game. One such example would be the Ohio State vs Michigan game, or as some call it, "The Game". A heavily betted contest, where lines are often established well before the week of the actual game.

3.4 Procedure

As said above, the data used in the analysis for the 2023 NCAA Division 1 college football season was collected from (CollegeFootballData.com, 2023). Of the 2,935 games, many of these games did not have the complete set of the betting lines described above. These games and their partial data were deleted from the overall data set. Once these games were deleted, the overall data set had 1,085 games. Although most betting lines have an added sense of bias since a large number bets come from fans and/or alumnus of the colleges in the game, so all school names were omitted to reduce any such bias. Next, the main procedure was to dichotomize the each of the variables. For DichoSpreadCovered, 1 if the favored team covered and 0 if not. DichoFavoredHome, 1 if the favored team won the game and 0 if they lost or the visiting team won. DichoChangeSpread was labeled as 1 if there was any change, increase or decrease, in the spread from opening to game time and see if there was no change. DichoChangeOU, is similar to the previous variable but with regards to the over/under line changing between opening and closing. However, each of the home and away money lines were not coded dichotomously. A dichotomous coded money line would be too challenging since the lines are as high as 50,000 compared to a change in spread that could be as high as 10. Also, there was no need for any IRB application as this data and study does not study human interactions, but rather the score of a collegiate football game.

3.5 Data Processing and Analysis

For the statistical analysis of this specific study of the predicting betting lines of the 2023 NCAA collegiate football season, all analyses were conducted using R (R Core Team, 2024). The predicted variable is whether or not the favored team covers the spread and so logistical model was used. Specifically, the package "CAR", Companion to Applied Regression, was used. The package offers a plethora and comprehensive techniques to analyze general logistic regression models (GLM). In addition to the robust GLM diagnostics that will be shown and used in the subsequent chapter four, the companion to applied regression, CAR, package enhances the variance inflation factor, VIFs, in order to detect multi-collinearity in the independent variables.

Determining fairness of bet is not a new concept but rather is an idea that has been around for decades and applies to numerous other sports. An example of prior related research, is an article that tests the fairness of spread betting on underdogs in the NFL (Badarinathi & Kochman, 1996). Although the article is testing fairness of spread betting, it gathers its data from the NFL and not college football. However, there is another article that indeed tests the fairness of spread betting in college football (Paul et al., 2003). Although the focus of the article was to determine the profitability of betting on home underdogs, the authors still tested for the fairness of those types of bets thereby justifying the idea of doing so for the 2023 college football season. Finally, an article that uses heavy machine learning to predict the outcome of a soccer match (Guan & Wang, 2022). The main difference between the aforementioned article and this paper presented is the type of machine learning being used. The machine learning in this paper uses a pre-written command from a publicly available R package, whereas in the articles just mentioned, the authors developed their own machine learning code to analyze results and predict outcomes. All three of these articles were dissected in depth in the literature review presented in chapter 2. A combination of these two previous articles justify the usage of the statistical model and hypothesis of this paper through their research using heavy machine learning to predict sport match outcomes and fairness of spread betting in college football.

3.6 Summary

This paper seeks to test the fairness of betting on the favored team to cover the spread for the 2023 college football season. In particular, it only considers games that involve a Division I FBS team. The necessity of such a study lies in the novelty of the approach when conducting the experiment. The use of machine learning in the linear regression model pioneers its away into the sports betting market. As discussed in the literature review section testing the fairness of spread betting has been around for decades and researchers often conduct experiments in their desired sport. However, testing the fairness of over/under betting and money line betting is also a relatively new subject. Furthermore, continuously testing the fairness of any of the aforementioned betting lines is necessary to ensure the integrity of bookmakers and the betting lines they are producing.

4 Results

4.1 Introduction

Presented in this chapter are the statistical result to the main research question. In addition to the main research question there are several hypotheses posed to give a more in-depth understanding and analysis to the main question: Are betting lines accurate predictors of spread coverage, over/under results and other related game outcomes. This will be answered by a series of smaller and more intricate questions posed as the hypotheses listed below.

The data gathered for this study allowed for versatility in research questions and hypotheses posed. It should be noted that data cleaning was discussed in the previous chapter. Certain games were removed from the data set due to missing components which made them unmeasurable. For any further detail on the removal of games from the data set see Chapter 3. Overall, the purpose of this chapter present the results of the analysis conducted on the data.

Additionally, this chapter will explore the implications of these findings for future betting strategies and predictive models. The conclusions drawn aim to contribute valuable insights to the field of sports analytics.

4.2 Data Descriptives and Analysis

4.2.1 Descriptives

A total of 1,073 games from the 2023 NCAA College Football Division I regular season were taken as the data set for this study. All of these games will serve as the sample size and any college football game played will be considered the population for the purpose of this study. In general it is found that 0.6281 or about 62.81% of favored teams covered their spread, while only 0.3719 or 37.19% did not cover. It was also found that in 0.5098 or 50.98% of games the over was made, while 0.4902 or 49.02% the under was made. Neither the home money lines nor away money lines served as accurate predictors of favored teams covering the spread.

These findings suggest that while favored teams tend to cover the spread more often than not, betting lines do not consistently predict spread coverage. Furthermore, the near-even distribution of over/under results indicates the inherent unpredictability of game outcomes.

We also see that teams who do cover the spread do so on average by a margin of 16.33 points. Similarly, teams who do not cover the spread fall short by an average margin of 7.77 points. Games whose total exceeded the final over/under did so on average by 14.11 points. Conversely, for game totals who fell short of the final over/under did so on average by a margin of 11.98 points. Another interesting statistic found was that the winning team on average won by a margin of 15.6 points.These statistics highlight the variability in game outcomes and provide insight into the extent to which teams exceed or fall short of expectations. Understanding these margins can be useful for developing more accurate predictive models and improving betting strategies.

4.2.2 Data Analysis

Presented below are both scatter plots and histograms for each of the spread and over/under changes within the entire sample. Each graph shows a very even distribution around the "no change" line, indicating a mostly zero change from opening to closing lines. These visuals emphasize the stability of betting lines over time, suggesting that the initial predictions remain consistent throughout a given time. This "mostly" even distribution also reflects the balanced nature of the betting responses to new team information as it becomes available.

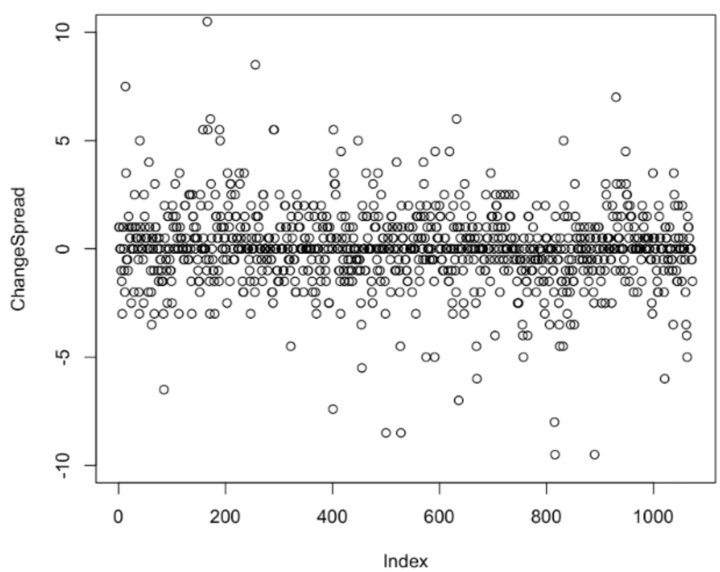

Figure 1: Numerical Change in Spread

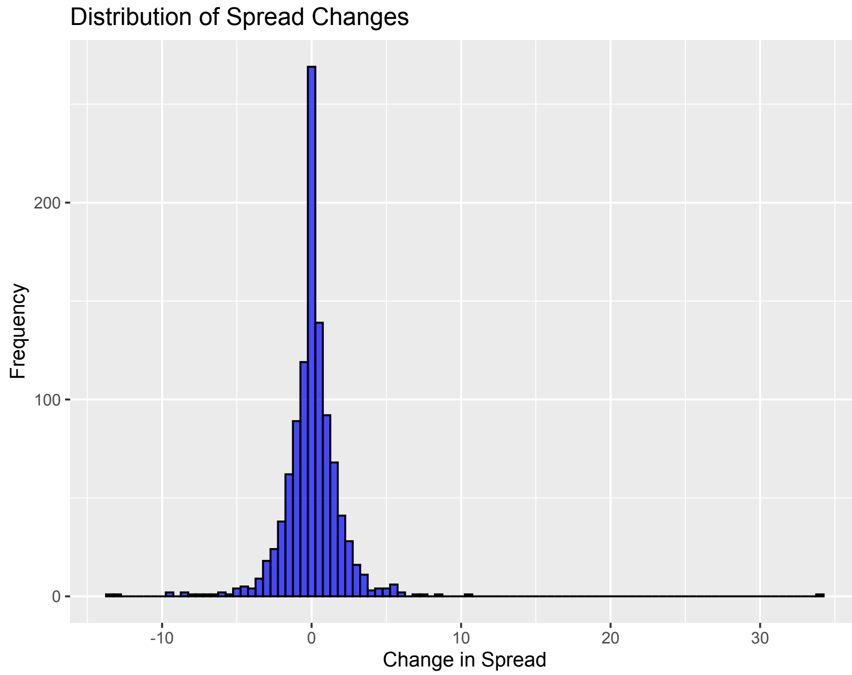

Figure 2: Distribution of Change in Spread

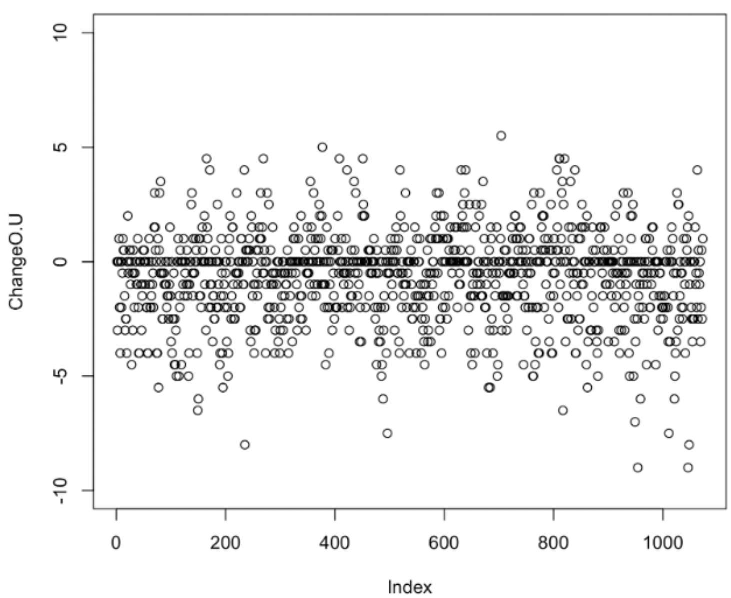

Figure 3: Numerical Change in Over/Under

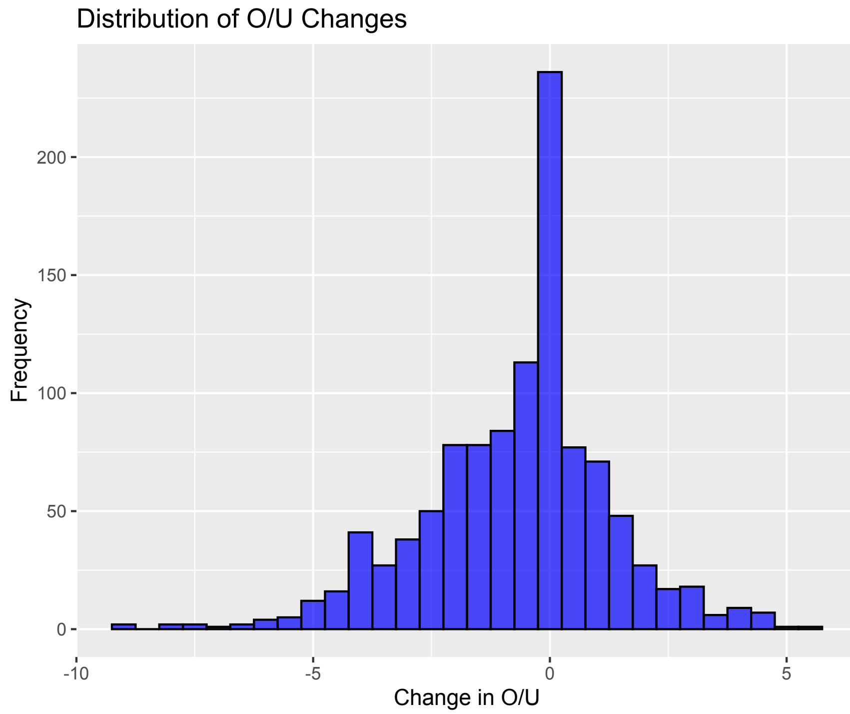

Figure 4: Distribution of Change in Over/Under

For both the spread and over/under, the only data collected was the opening and closing lines. By the time this thesis was started the 2023 season had already been completed. Consequently, it was not possible to track movement during each week. Tracking betting lines as the move to observed trends (see the example in figure 3) could serve as a future research topic. Previous literature suggests that games that do not draw media attention, thus causing the betting lines to remain stagnant, are often inaccurate. Therefore, betting against such lines could be profitable.

4.3 Hypotheses

The main hypothesis: Changes in spread, over/under and the home team being the favored team are accurate predictors of the favored team covering the final spread. The hypothesis posed serves as a general overview for the research being conducted in this thesis and has consequently been broken down into seven sub-hypotheses to provide a more robust analysis of the data collected. The theory behind the breakdown, which was discussed in Chapter 2 in greater detail, has to do with bookkeepers and their attempt to keep total bets split 50/50 to ensure their own profits. Since those same bookkeepers handle spread, over/under and other betting lines it seems logical to test statistical relationship between them. The results will be synthesized together to draw the main conclusion.

Logistic Regression was used to conduct the analysis across all of the sub-hypotheses for consistency as well as for optimal evaluation. Although, in hypotheses 4 and 5 it was necessary to perform an AIC and BIC test to assess whether closing lines or opening lines are better predictors of covering the spread and taking the over. Naturally, the data set used by each of the seven hypotheses remains the same. Each sub-hypothesis uses betting lines as predictors on another betting line being met. The two dependent variables used were DichoSpreadCovered and DichoO.U. DichoSpreadCovered contains 1 if the favored team covered the spread and 0 if they did not. DichoO.U contained a 1 if it was consider over and a 0 if it was considered under. ChangeSpread contains the numerical value for the change in spread which could either be positive (favored team must win by less) or negative (favored team must win by more). ChangeO.U contains the numerical values of the change between opening and closing under/under lines. DichoChangeSpread contains a 1 if there was any change between opening and closing spread and 0 if there was no change. DichoFavored.Home contains a 1 if the favored team is the home team and 0 if the favored team is the away team. DichoChangeO.U contains a 1 if there was any change between opening and closing over/under and 0 if there was no change. Finally, the last 4 independent variables are trivial: opening and closing spread, and opening and closing over/under.

4.4 Hypothesis 1

The first hypothesis is as follows: Changes in spread, over/under and the home team being the favored team are accurate predictors of the favored team covering the final spread.

Hypothesis 1 Coefficients				
	Estimate	Std. Error	z Value	Pr(>|z|)
Intercept	-0.1337	0.1024	-1.3050	0.1920
DichoFavored.Home	0.0156	0.1267	0.1230	0.9020
ChangeSpread	0.0322	0.0296	1.0860	0.2770
ChangeO.U	0.0198	0.0313	0.6340	0.5260

Table 1: Hypothesis 1 Coefficients

Above is the display of the logistical regression model and the relevant coefficients for Hypothesis 1.

The table indicates that there was no statistical significance at the 0.05 level for any of the variables tested. These results are confirmed based on the confidence intervals below. Since each interval contains zero there is no statistical significance.

Hypothesis 1 Confidence Intervals		
	2.5 %	97.5 %
Intercept	-0.3350	0.0668
DichoFavored.Home	-0.2330	0.2640
ChangeSpread	-0.0248	0.0297
ChangeO.U	-0.0414	0.0813

Table 2: Hypothesis 1 Confidence Intervals

Upon examining the logistic regression model summary of hypothesis one, the observed p-values for the predictor variables are high, indicating that they are not statistically significant at the 0.05 level. This lack of significance suggests that our predictors do not contribute meaningful information to the model for predicting the favored team covering the spread. The residual vs. fitted graph below supports this finding, as it shows two distinct lines of residuals around plus and minus one. This pattern is typical in logistic regression due to the dichotomy of the dependent variable, the favored team covering the spread. However, the clustering of the residuals at these values suggests that the model predictions are too extreme (close to 0 or 1) and do not vary sufficiently.

The data clustering around +1 (the team covered but the model predicted the probability of covering was closer 0) and -1 (the team did not cover but the model predicted the probability of covering was closer to 1) indicate that the model's predictions are extreme and often incorrect. This indicates that our predictors are not effectively differentiating between the two outcome classes of the dependent variable, leading to poor model performance. Together, these findings suggest our model was insufficient in predicting the favored team covering using if the favored team is also the home team, if there was any change in the spread and/or over/under lines.

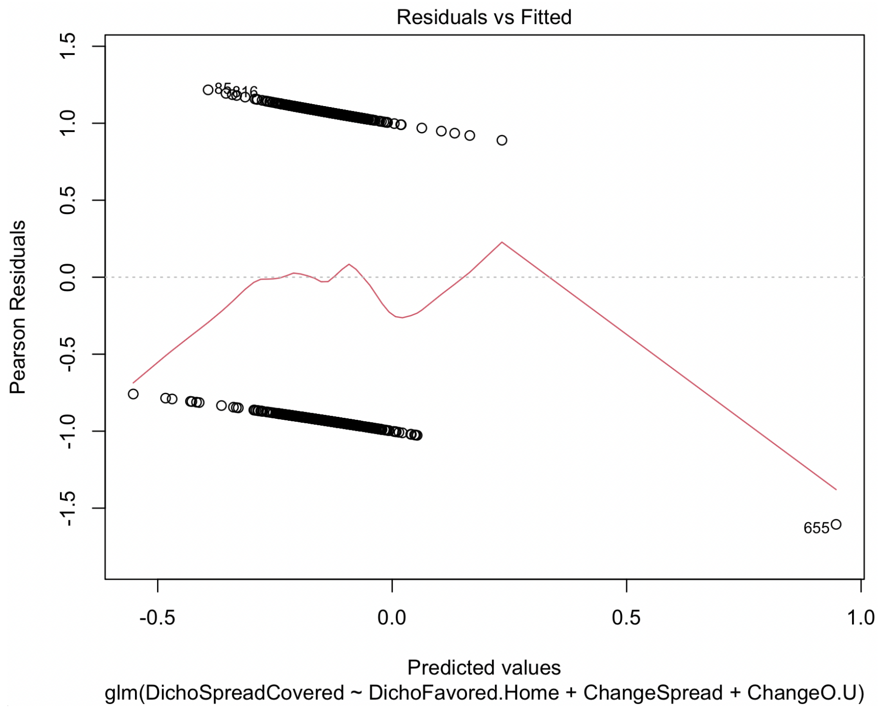

Figure 5: Hypothesis 1 Residuals

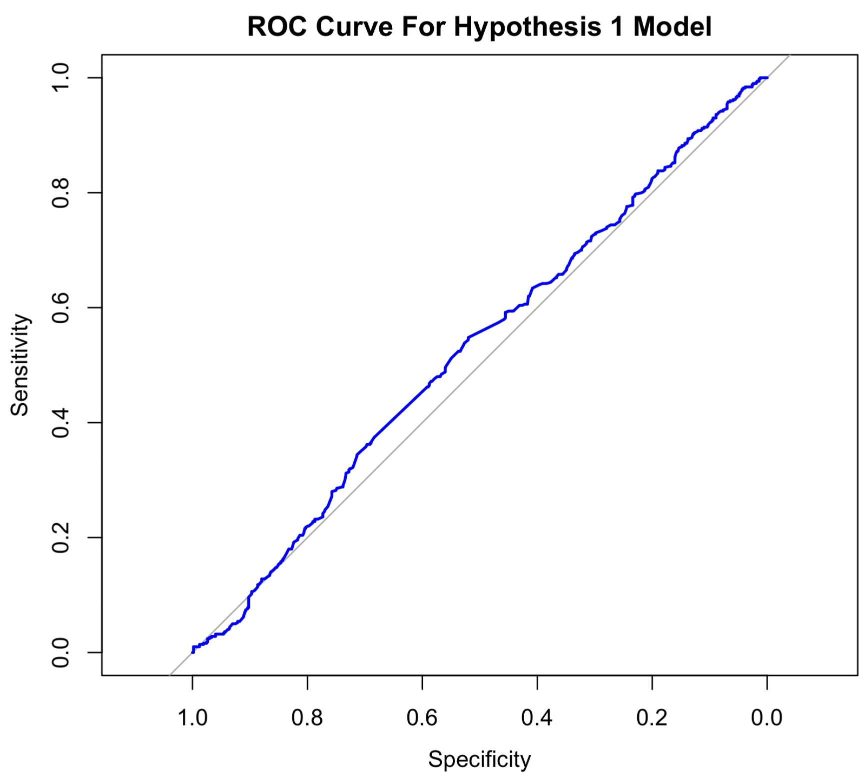

Figure 6: Hypothesis 1 ROC Curve

Additionally, when analyzing the ROC curve for the model, predicting whether a team covers the spread, lacks discriminating ability. This implies that the model is almost as likely to incorrectly classify whether a team will cover the spread (ie. not to cover spread) as it is to correctly classify it (ie. cover the spread). Essentially, the model's predictions are no better than a random guess of if the favored team will cover the spread.

4.4.1 Hypothesis 2

The second hypothesis tested is as follows: Games with having any changes between the opening and closing spreads are more likely to accurately predict if the favored team will cover the spread.

Hypothesis 2 Coefficients				
	Estimate	Std. Error	z Value	Pr(>|z|)
Intercept	-0.1373	0.0612	-2.2420	0.0250*
ChangeSpread	0.0200	0.0293	1.0210	0.3070

Table 3: Hypothesis 2 Coefficients

Above is the display of the logistic regression model's relevant coefficients for Hypothesis 2. The table indicates that there is statistical significance at the 0.05 level with the intercept, but not with the ChangeSpread independent variable. This implies that, in general, the favored team does cover the spread regardless of the change between the opening and closing spreads. However, the main focus of the hypothesis is confirmed below since the interval for the independent variable contains 0, and thus there is no statistical significance.

Hypothesis 2 Confidence Intervals		
	2.5 %	97.5 %
Intercept	-0.2575	0.0174
ChangeSpread	-0.0266	0.0898

Table 4: Hypothesis 2 Confidence Intervals

4.4.2 Hypothesis 3

The third hypothesis is as follows: Positive changes between opening and closing spreads are more likely to predict if the favored team will cover the spread. A new data frame was created using R to filter out the rows with only positive changes in the spread. The new frame had an index of 419 data points.

Hypothesis 3 Coefficients				
	Estimate	Std. Error	z Value	Pr(>|z|)
Intercept	0.1093	0.1331	0.8220	0.4110
ChangeSpread	-0.0606	0.0586	-1.0350	0.3010

Table 5: Positive Changes in Spread

Above is the display of the logistical regression model's relevant coefficients for Hypothesis 3. The table indicates that there was not statistical significance at the 0.05 level for the of the variable tested. This result is confirmed below based on the confidence interval. Since our variable's confidence interval contains zero, there is no statistical significance.

Hypothesis 3 Confidence Intervals		
	2.5 %	97.5 %
Intercept	-0.1418	0.3835
ChangeSpread	-0.1941	0.0401

Table 6: Hypothesis 3 Confidence Intervals

4.4.3 Hypothesis 4

The fourth hypothesis is as follows: Negative changes between opening and closing spreads are likely to predict if the favored team will cover the spread. As in Hypothesis 2, a new data frame was used with only rows sharing a negative change in the spread. This frame's index had 386 data points.

Hypothesis 4 Coefficients				
	Estimate	Std. Error	z Value	Pr(>|z|)
Intercept	-0.2920	0.1453	-2.0090	0.0445*
ChangeSpread	0.0197	0.0631	0.3120	0.7548

Table 7: Negative Changes in Spread

Above is the display of the logistical regression model's relevant coefficients for Hypothesis 4. The table indicates that there is statistical significance at the 0.05 level with the Intercept but not with the ChangeSpread independent variable which implies that in general the favored team does cover the spread regardless of the change between opening and closing spread. However, the main focus of the hypothesis is confirmed below since the interval for the independent variable contains 0 and thus there there is no statistical significance.

Hypothesis 4 Confidence Intervals		
	2.5 %	97.5 %
Intercept	-0.5773	0.0062
ChangeSpread	-0.1023	0.1485

Table 8: Hypothesis 4 Confidence Intervals

The probabilities in favor of covering the spread for each positive changes and negative changes in the spread were found using the two previous models. As there was 419 and 386 data points respectively, the mean of their probabilities was found. The model only containing only the positive changes had a mean probability of about 0.5036 or 50.36%. On the negative side, the mean probability of a negative change in the spread was about 0.4197 or 41.97%. A third additional model for only the games with no change, which totaled 268 points, was created. But as there was no variance in the data points, the model's descriptive statistics were uninteresting and so were not shown above. However, the model was able to predict the probabilities of the games with no change in the spread, covering the spread. The mean of those probabilities was about 0.4739 or 47.39%. This is shown below:

Positive Changes in Spread

- **Data Points**: 419

- **Mean Probability of Covering the Spread**: 0.5036 (50.36%)

Negative Changes in Spread

- **Data Points**: 386
- **Mean Probability of Covering the Spread**: 0.4197 (41.97%)

No Changes in Spread

- **Data Points**: 268
- **Mean Probability of Covering Spread**: 0.4739 (47.39%)
- **Model Variance**: Due to no variance in the predictor (spread change), the model's descriptive statistics were not informative.

The models for positive and negative changes were able to provide meaningful insights into the probability of covering the spread based on the direction of the spread change. The additional model for games with no change in the spread showed a mean probability but lacked variability to yield interesting descriptive statistics.

4.4.4 Hypothesis 5

The fifth hypothesis is as follows: Closing over/under lines are more accurate predictors of game totals than opening over/under lines.

Hypothesis 5 Coefficients				
	Estimate	Std. Error	z Value	Pr(>|z|)
Intercept	0.7589	0.4389	1.729	0.0838
OpeningOU	-0.0138	0.0329	-0.968	0.3332
ClosingOU	0.0184	0.0311	0.591	0.5543

Table 9: Hypothesis 5 Coefficients

Above is the display of the logistical regression model's relevant coefficients for Hypothesis 5. The table indicates that there was not statistical significance at the 0.05 level for any of the variables tested. This results is confirmed based on the confidence interval below. Since each interval contains zero there is no statistical significance.

Hypothesis 5 Confidence Intervals		
	2.5 %	97.5 %
Intercept	-0.5773	1.6230
OpeningOU	-0.1023	0.0326
ClosingOU	-0.0425	0.0795

Table 10: Hypothesis 5 Confidence Intervals

For hypothesis #5 the dependent variable selected is total game score. The hypothesis was tested as three different models followed by a AIC and BIC comparison to determine which model is the better fit. Table 9 uses both opening and closing over/under as the independent variables. Comparatively, table 11 has only opening over/under as an independent variable.

Opening O/U				
	Estimate	Std. Error	z Value	Pr(>|z|)
Intercept	0.7232	0.4346	1.664	0.0961
OpeningOU	0.0110	0.0082	-1.590	0.1118

Table 11: Opening Over/Under

Similarly, Table 12 has only closing over/under as an independent variable. Looking at the tables presented for Hypothesis 5 thus far, the results do not indicate their is statistical significance for the independent variables used.

Closing O/U				
	Estimate	Std. Error	z Value	Pr(>|z|)
Intercept	0.5990	0.4062	1.475	0.140
ClosingOU	-0.0108	0.0077	-1.394	0.163

Table 12: Closing Over/Under

Below is a comparison of the AIC and BIC results of the three models. The first uses OpeningOU as its only independent variable and the second uses ClosingOU as its only independent variable. However, the third model uses both opening and closing over/under as independent variables which is represented in the table as “Opening & Closing”. The results below indicate that the model only using opening over/under is a better for the data.

Over/Under Comparatives			
	df	AIC	BIC
OpeningOU	3	1488.545	1498.502
ClosingOU	3	1489.133	1499.090
Opening&Closing	3	1490.196	1505.130

Table 13: Hypothesis 5 AIC & BIC Comparatives

4.4.5 Hypothesis 6

The sixth hypothesis is as follows: Closing spread lines are more accurate predictors of spread coverage than opening spread lines.

Hypothesis 6 Coefficients				
	Estimate	Std. Error	z Value	Pr(>|z|)
Intercept	-0.0693	0.1045	-0.663	0.507
OpeningSpread	-0.0244	0.0300	-0.812	0.417
ClosingSpread	0.0310	0.0292	1.060	0.289

Table 14: Hypothesis 6 Coefficients

Above is the display of the logistical regression model’s relevant coefficients for Hypothesis 6. The table indicates that there was not statistical significance at the 0.05 level for any of the variables tested. This results is confirmed based on the confidence interval below. Since each interval contains zero there is no statistical significance.

Hypothesis 6 Confidence Intervals		
	2.5 %	97.5 %
Intercept	-0.2745	0.1356
OpeningSpread	-0.0854	0.0337
ClosingSpread	-0.0256	0.0904

Table 15: Hypothesis 6 Confidence Intervals

For hypothesis 6, the dependent variable selected is DichoSpreadCovered. The hypothesis was tested as three different models followed by a AIC and BIC comparison to determine which model is the better fit. Table 14 uses both opening and closing spreads as the independent variables. Comparatively, table 16 has only OpeningSpread as an independent variable.

Opening Spread				
	Estimate	Std. Error	z Value	Pr(>|z|)
Intercept	-0.0726	0.1044	-0.695	0.487
OpeningSpread	0.0062	0.0082	0.751	0.452

Table 16: Opening Spread

Similarly, Table 17 has only closing spread as an independent variable. Looking at the tables presented for Hypothesis 6 thus far, the results do not indicate their is statistical significance for the independent variables used.

Closing Spread				
	Estimate	Std. Error	z Value	Pr(>|z|)
Intercept	-0.0523	0.1025	-0.511	0.609
ClosingSpread	0.0082	0.0080	1.019	0.308

Table 17: Closing Spread

Below is a comparison of the AIC and BIC results of the three models. The first uses OpeningSpread as its only independent variable and the second uses ClosingSpread as its only independent variable. However, the third model uses both opening and closing spreads as independent variables which is represented in the table as “Opening & Closing”. The results below indicate that the model only using closing over/under is a better for the data.

Spread Comparatives			
	df	AIC	BIC
OpeningSpread	3	1486.809	1495.914
ClosingSpread	3	1485.958	1495.438
Opening & Closing	3	1485.481	1501.744

Table 18: Hypothesis 6 AIC & BIC Comparatives

4.4.6 Hypothesis 7

The seventh hypothesis is as follows: Changes in spread, over/under, and the home team being the favored team are accurate predictors of the total game points related to the closing over/under.

Hypothesis 7 Coefficients				
	Estimate	Std. Error	z Value	Pr(>|z|)
Intercept	0.2175	0.1173	-1.854	0.0637
DichoFavoredHome	-0.1667	0.1268	-1.314	0.1888
ChangeSpread	0.0528	0.0308	1.714	0.0865
ChangeOU	-0.0504	0.0417	-1.207	0.2272

Table 19: Hypothesis 7 Coefficients

Above is the display of the logistical regression model’s relevant coefficients for Hypothesis 7. The table indicates that there was not statistical significance at the 0.05 level for any of the variables tested. This results is confirmed based on the confidence interval below. Since each interval contains zero there is no statistical significance.

Hypothesis 7 Confidence Intervals		
	2.5 %	97.5 %
Intercept	-0.0119	0.4482
DichoFavoredHome	-0.4156	0.0182
ChangeSpread	-0.0054	0.1155
ChangeOU	-0.1326	0.0313

Table 20: Hypothesis 7 Confidence Intervals

Hypothesis 7 uses the same independent variables but has DichoO.U as the dependent variable and thus serves as an extension of Hypothesis 1. It feels natural to include the over/under as the dependent variable since it is one of the two main betting lines being study in this thesis. The ROC curve, as shown below, remains remarkably similar to ROC curve for Hypothesis 1 and therefore the conclusions remain the same despite some minor changes to accommodate the change in dependent variables.

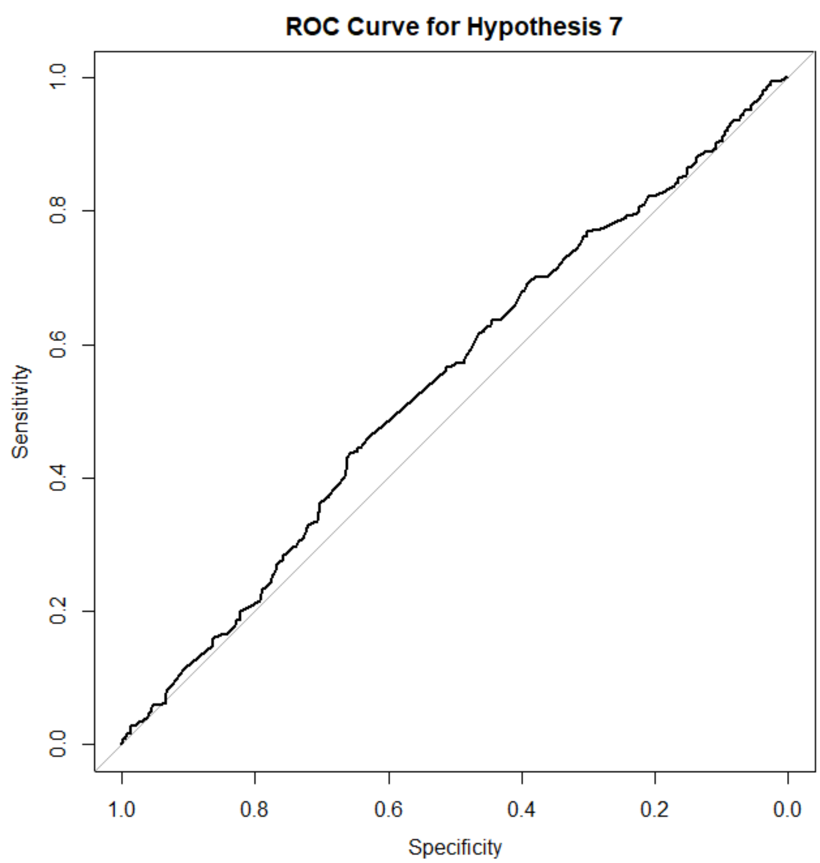

Figure 7: Hypothesis 7 ROC Curve

4.5 Conclusion

The overall results presented in this chapter have significant implications in the field of sports gambling. Each of the hypotheses above went through significant discussion and revisions to ensure the experiments had broad implications while maintaining intricacy. To do so, certain variables were dropped or excluded from the study as it progressed due to difficulties in implementation or irrelevancy in theory. Despite these changes the final seven hypotheses presented above offer a dynamic analysis of the data. In the concluding subsequent chapter this thesis will offer connections to other current research, how these experiments could evolve into further research, as well as where the field will be going in the near future. Additionally, limitations of the experiments conducted for this thesis will also be discussed once more.

5 Conclusions

5.1 Interpretation of Findings

Each model provided a unique insight into the analysis of the data. The first hypothesis states that the numerical change in spread, numerical change in over/under, and the favored team being the home team are significant predictors of the favored team covering the spread. However, the statistical results from the linear regression model concludes that there is no evidence to support this hypothesis and as such the null hypothesis is rejected. A point-biserial correlation test was conducted between the dichotomous variable (DichoFavoredHome) and each continuous independent variable (ChangeSpread and ChangeOU), confirming that the assumption of independence was maintained. Furthermore, the Variation Inflation Factor (VIF) was calculated for the model and concludes that there are no signs of multicollinearity.

The second hypothesis states that games with changes, positive or negative, between opening and closing spread are accurate predictors if the favored team covers the spread. The lone independent variable, ChangeSpread, holds the numerical changes between opening spread and closing spread. For example, if the opening spread was -5 and closing spread was -4 then the number stored in the variable would be 1. A positive change in spread means the favored team must now win by less points than what was specified at the opening spread. Conversely, a change from -5 to -6 means a -1 will be stored in the variable. A negative change indicates that the favored team must now win by more than originally predicted. As shown in both the coefficients and the confidence interval, no statistical significance was found and therefore the null hypothesis is rejected. In other words, there is not enough evidence to conclude that ChangeSpread is an accurate predictor for if the favored team will cover the spread. The point-biserial correlation test results are applicable to hypothesis 2 since the test is individual variables and not the unique model. A VIF was not caclulated for this model since it contained only one predictor and therefore multicollinearity is impossible.

The third hypothesis states that games with positive changes between opening and closing spreads are accurate predictors if the favored team will cover the spread. As explained in the summary of the second hypothesis, a positive change between opening and closing spreads indicates the favored team must now win by less points than originally predicted. The data frame was reduced from 1,073 to 419 since not every game had a change in spread. Additionally, since the model contained only one independent variable a VIF was not calculated and thereby eliminating threats of multicollinearity. With both the confidence interval and coefficients indicating no statistical significance therefore it cannot be said that positive changes in over/under are accurate predictors of the favored team covering the spread. A VIF was not calculated for the model since only one independent variable was used and thus concludes that multicollinearity is a non-issue.

The fourth hypothesis states that games with negative changes between opening and closing spreads are accurate predictors if the favored team will cover the spread. This hypothesis was created as a follow up to the previous hypothesis since logically testing both the negative and positve changes for significance is plausible. As explained in the summary of the second hypothesis, a negative change between opening and closing spreads indicates the favored team must now win by more points than originally predicted. The data frame was reduced from 1,073 to 386 since not every game had a change in spread. Since the model also contained only one independent variable a VIF was not calculated and thereby eliminating threats of multicollinearity for this model. With both the confidence interval and coefficients indicating no statistical significance therefore it cannot be said that changes in over/under are accurate predictors of the favored team covering the spread. A VIF was not calculated for the model since only one independent variable was used and thus concludes that multicollinearity is a non-issue.

The fifth hypothesis states that closing over/under lines are stronger and more accurate predictors of the spread coverage than opening over/under lines. OpeningOU/ClosingOU contains the numerical values of the opening over/under and closing over/under. This hypothesis tested three different models using one dependent variable and two different independent variables. The first model tested opening over/under lines as a predictor for spread coverage and did not see any statistical significance. The second model tested closing over/under lines as a predictor and also did not see any statistical significance. Thirdly, a model was created using both the opening over/under and closing over/under as the independent variables and again no statistical significance was found. However, a VIF was calculated on the third model and found there to be issues with multicollinearity. Independence was maintained across all three models and a VIF was not able to calculated on the first two since they each only used one independent variable. To determine which model is a better fit the three models were compared to one another using their Bayesian information criterion (BIC) and Akaike information criterion (AIC). The comparison revealed that the model using only the opening over/under lines was a better for the data.

The sixth hypothesis states that closing spread lines are stronger and more accurate predictors of the spread coverage than opening spread lines. This hypothesis follows a similar premise to hypothesis 5, except instead of OpeningOU/ClosingOU, it uses OpeningSpread/ClosingSpread.OpeningSpread/ClosingSpread contains the numerical values of the opening spread and closing spread. This hypothesis tested three different models using one dependent variable and two different independent variables. The first model tested opening spread lines as a predictor for spread coverage and did not see any statistical significance. The second model tested closing spread lines as a predictor and also did not see any statistical significance. Thirdly, a model was created using both the opening spread and closing spread as the independent variables and again no statistical significance was found. However, a VIF was calculated on the third model and found there to be issues with multicollinearity. Independence was maintained across all three models and a VIF was not able to calculated on the first two since they each only used one independent variable. To determine which model is a better fit the three models were compared to one another using their Bayesian information criterion (BIC) and Akaike information criterion (AIC). The comparison revealed that the model using only the closing spread lines was a better for the data.

Hypothesis 7 is almost identical to Hypothesis 1 with the exception of the dependent variables. Rather than using the numerical change in spread, the numerical change in over/under, and whether the favored team is the home team to predict if the favored team covered the spread, these same independent variables are used to predict if the total game score was over or under the final betting line. The dependent variable is dichotomously coded with 0 if the game score fell under and 1 if the game score went over the final predicted total game score. Similar the previous six hypotheses, there was no statistical significance found by the logistic regression ran on the model as indicated by both the confidence interval and the coefficients. Independence between the three variables used in this model was maintained. Finally, a VIF was calculated and determined there are no threats of multicollinearity.

5.2 Connections to Existing Literature

Our research into the relationship between the change in college football betting lines for the spread and over/under lines and whether the favored team covers the spread directly ties into the efficient market hypothesis (EMH). Pankoff showed that the NFL betting market was "found to be efficient in the aggregate." (Pankoff, 1968). According to EMH, in the context of sports betting, this would mean that the odds and spreads set by bookmakers are based on all available information about teams, players, and countless other factors, making it difficult for bettors to gain an edge by analyzing changes in spreads. If our analysis finds that changes in spreads can predict whether a favored team covers the spread, it could suggest inefficiencies in the betting market, challenging the EMH.

Conversely, if our research shows no statistical significant between the change in college football betting lines for the spread and over/under lines and whether the favored team covers the spread, it would support the notion that sports betting markets are efficient. Bookmakers may and will adjust these said spreads to incorporate new information, and bettors cannot consistently exploit these adjustments to gain an advantage. This would align with the EMH, suggesting that the spreads reflect all known information and that any attempt to predict outcomes based on spread changes would be comparable to attempting to beat a well-functioning stock market. Therefore, our research not only provides insights into betting strategies but also contributes to the broader understanding of market efficiency in the context of sports betting.

5.3 Recommendations for Future Research

Moving forward there are plenty of ways for other researchers to build upon this study and publish their own findings. One such example would be using machine learning or artificial intelligence (AI) to run similar types of models to those presented here. A prime example of this is an article that was discussed in Chapter 1 (Guan & Wang, 2022). Using machine learning to create a more complex model might lead to the discovery of new predictors for spread coverage in a way that this thesis could not. Aside from spread coverage, the ultimate objective could now shift towards accurately predicting game winners and total scores. being able to accurately just one of the three would be an extreme breakthrough. Furthermore, once obtaining such an accurate model, it is reasonable to take it one step farther and have AI/machine learning come up with profitability simulation to maximize betting profit.

Another potential avenue for future research is creating a profitability simulation based on these results. Creating such simulations was not the primary focus of this thesis but there is still ample opportunity for research in this area. It is common for articles to run these simulation despite minimal statistical significance appearing in the actual research (Paul et al., 2003). It is important to note that throughout this thesis, spread coverage is the main focus and not the actual outcome of the game. For example, just because the favored team won does not mean they covered the spread. Disregarding spread coverage, 74% of favored teams won their respective games which is a rather high percentage and deserves to be investigated further. Despite the lack of evidence to find predictors for spread coverage, it is possible to create a betting simulation that would maximize profit from betting. Creating these betting simulation to maximize profit require an intricate understanding of the data and significant knowledge in a coding language such as python. To focus on the main research questions proposed, the concept was left to the reader as an avenue for future research since implementation would require additional time and research.

5.4 Conclusion

In conclusion, the researched conducted for this thesis yields promising results for the field of sports betting. Although it may seem the lack of statistical significance implies for all of the hypotheses implies there are no applicable results for the field of sports betting, that is not the case. In fact it is because of the those results there are implications for the field. Specifically, predicting spread coverage is severely more difficult than expected and either requires more research or even impossible to do so. Given how many different aspects there are to a college football game it may seem impossible to test all of them as predictors. The theories and ideas discussed are greatly influenced and supported by previous literature while still maintaining originality. Making use of the 2023 NCAA Div I regular season had its challenges but still effectively served as a sample for the population of college football. Additionally, as previously stated, the goal of this paper is not encourage the abuse of gambling but readers are advised to do so at their own responsibility and discretion. Although profitability models were not discussed or presented in this thesis, it serves as an opportunity for the reader to conduct further research. Profitability models in conjunction with machine learning provide ample opportunities for growth in the field of sports betting. The reader is encourage to use the data set, previous literature and this thesis as tools for creating original research.

References

APNews. (2023, August). *Iowa st qb dekkers accused of betting on cyclones sports, charged with tampering in gambling probe.* Retrieved 2024-02-11, from https://apnews.com/article/dekkers-iowa-state-football-ncaa-7280feaf600febeef9b9f4e958a66a9b (Section: Sports)

Arscott, R. (2023, June). Market efficiency and censoring bias in college football gambling. *Journal of Sports Economics*, *24*(5), 664–689. Retrieved 2024-01-12, from http://login.proxy01.shawnee.edu/login?url=https://search.ebscohost.com/login.aspx?direct=true&db=s3h&AN=163453007&site=eds-live&scope=site (Publisher: Sage Publications Inc.)

Badarinathi, R., & Kochman, L. (1996, October). Football Betting and the Efficient Market Hypothesis. *The American Economist*, *40*(2), 52–55. Retrieved 2024-02-06, from http://journals.sagepub.com/doi/10.1177/056943459604000207 doi: 10.1177/056943459604000207

Bengel, C., McCarriston, S., & Staff, C. S. (2023). *U.s. sports betting: Here is where all 50 states currently stand on legalizing online sports betting sites - cbssports.com.* Retrieved 2024-02-06, from https://www.cbssports.com/general/news/u-s-sports-betting-here-is-where-all-50-states-currently-stand-on-legalizing-online-sports-betting-sites/

Berkowitz, J. P., Depken, C. A., & Gandar, J. M. (2018, October). The Conversion of Money Lines Into Win Probabilities. *Journal of Sports Economics*, *19*(7), 990–1015. Retrieved 2024-02-14, from http://login.proxy01.shawnee.edu/login?url=https://search.ebscohost.com/login.aspx?direct=true&db=s3h&AN=131574917&site=eds-live&scope=site (Publisher: Sage Publications Inc.)

Coleman, B. J. (2017, May). Team travel effects and the college football betting market. *Journal of Sports Economics*, *18*(4), 388–425. Retrieved 2024-02-06, from http://journals.sagepub.com/doi/10.1177/1527002515574514 doi: 10.1177/1527002515574514

Collegefootballdata.com. (2023). Retrieved 2024-01-12, from https://collegefootballdata.com/exporter/games?year=2022&seasonType=regular

Cox, J., Schwartz, A. L., Van Ness, B. F., & Van Ness, R. A. (2021, April). The predictive power of college football spreads: Regular season versus bowl games. *Journal of Sports Economics*, *22*(3), 251–273. Retrieved 2024-02-06, from http://login.proxy01.shawnee.edu/login?url=https://search.ebscohost.com/login.aspx?direct=true&db=s3h&AN=148704807&site=eds

-live&scope=site (Publisher: Sage Publications Inc.)

Dmochowski, J. P. (2023, June). A statistical theory of optimal decision-making in sports betting. *PLoS ONE*, *17*(6), 1–22. Retrieved 2024-02-06, from http://login.proxy01.shawnee.edu/login?url=https://search.ebscohost.com/login.aspx?direct=true&db=fsr&AN=164619600&site=eds-live&scope=site (Publisher: Public Library of Science) doi: 10.1371/journal.pone.0287601

Dodd, D. (2022, November). *College sports' overlooked gambling issue: Improprieties being ignored as betting soars nationwide.* Retrieved 2024-02-04, from https://www.cbssports.com/college-football/news/college-sports-overlooked-gambling-issue-improprieties-being-ignored-as-betting-soars-nationwide/

Gandar, J., Zuber, R., O'Brien, T., & Russo, B. (1988, September). Testing Rationality in the Point Spread Betting Market. *The Journal of Finance*, *43*(4), 995–1008. Retrieved 2024-02-06, from https://onlinelibrary.wiley.com/doi/10.1111/j.1540-6261.1988.tb02617.x doi: 10.1111/j.1540-6261.1988.tb02617.x

Guan, S., & Wang, X. (2022, February). Optimization analysis of football match prediction model based on neural network. *Neural Computing & Applications*, *34*(4), 2525–2541. Retrieved 2024-02-06, from http://login.proxy01.shawnee.edu/login?url=https://search.ebscohost.com/login.aspx?direct=true&db=iih&AN=155105257&site=eds-live&scope=site (Publisher: Springer Nature) doi: 10.1007/s00521-021-05930-x

Hodges, S., Lin, H., & Liu, L. (2013, March). Fixed Odds Bookmaking with Stochastic Betting Demands: European Financial Management. *European Financial Management*, *19*(2), 399–417. Retrieved 2024-03-17, from http://login.proxy01.shawnee.edu/login?url=https://search.ebscohost.com/login.aspx?direct=true&db=bth&AN=85922074&site=eds-live&scope=site (Publisher: Wiley-Blackwell) doi: 10.1111/j.1468-036X.2012.00601.x

Howington, E. B., & Moates, K. N. (2017, September). Is there a bye week advantage in college football?: Electronic Journal of Applied Statistical Analysis. *Electronic Journal of Applied Statistical Analysis*, *10*(3), 735–744. Retrieved 2024-03-18, from http://login.proxy01.shawnee.edu/login?url=https://search.ebscohost.com/login.aspx?direct=true&db=aph&AN=126532421&site=eds-live&scope=site (Publisher: Electronic Journal of Applied Statistical Analysis (EJASA)) doi: 10.1285/i20705948v10n3p735

Humphreys, B. R. (2010). Point spread shading and behavioral biases in nba betting markets. *Revista Di Diritto Ed Economia Dello Sport (RDES)*.

Krieger, K., & Davis, J. (2023, December). Examining the impact of visibility on market efficiency: lessons from movement in NFL betting lines: Journal of Economics & Finance. *Journal of Economics & Finance*, 1–17. Retrieved 2024-03-18, from http://login.proxy01.shawnee.edu/login?url=https://search.ebscohost.com/login.aspx?direct=true&db=bth&AN=174151397&site=eds-live&scope=site (Publisher: Springer Nature) doi: 10.1007/s12197-023-09656-5

Krieger, K., & Davis, J. L. (2022, January). How Much of "Home Field Advantage" Comes from the Fans? A Natural Experiment from the COVID-19 Pandemic: UNLV Gaming Research & Review Journal. *UNLV Gaming Research & Review Journal*, *26*(1), 59–72. Retrieved 2024-03-23, from http://login.proxy01.shawnee.edu/login?url=https://search.ebscohost.com/login.aspx?direct=true&db=bth&AN=161051046&site=eds-live&scope=site (Publisher: University of Nevada, Las Vegas)

Kuester, D. D., & Sanders, S. (2011, January). Regional information and market efficiency: the case of spread betting in united states college football. *Journal of Economics and Finance*, *35*(1), 116–122. Retrieved 2024-02-06, from http://link.springer.com/10.1007/s12197-009-9113-3 doi: 10.1007/s12197-009-9113-3

Levitt, S. D. (2004, April). Why are Gambling Markets Organised so Differently from Financial Markets? *The Economic Journal*, *114*(495), 223–246. Retrieved 2024-03-28, from https://academic.oup.com/ej/article/114/495/223-246/5086012 doi: 10.1111/j.1468-0297.2004.00207.x

Manuel, D. (2022, June). *There Are 834 College Football Teams In The United States As Of This Moment.* Retrieved 2024-03-28, from https://www.sports-king.com/how-many-college-football-teams-3338/

Meyer, C. (2024, January). *Michigan football sign-stealing scandal, explained: Why coach jim harbaugh was suspended.* Retrieved 2024-02-11, from https://www.freep.com/story/sports/college/university-michigan/2024/01/08/michigan-sign-stealing-scandal-jim-harbaugh-suspended-connor-stalions-big-ten-ncaa/72134775007/

Moore, E., & Francisco, J. (2019, June). Efficiency, Profitability, and College Football: do Bettors Overrate the Major Conferences in Interconference Play? *Atlantic Economic Journal*, *47*(2), 159–167.

Retrieved 2024-02-06, from http://link.springer.com/10.1007/s11293-019-09616-7 doi: 10.1007/s11293-019-09616-7

Pankoff, L. D. (1968). Market efficiency and football betting. *The Journal of Business*, *41*. Retrieved from https://EconPapers.repec.org/RePEc:ucp:jnlbus:v:41:y:1968:p:203

Paul, R. J., & Weinbach, A. P. (2008, August). Price Setting in the NBA Gambling Market: Tests of the Levitt Model of Sportsbook Behavior: International Journal of Sport Finance. *International Journal of Sport Finance*, *3*(3), 137–145. Retrieved 2024-03-28, from http://login.proxy01.shawnee.edu/login?url=https://search.ebscohost.com/login.aspx?direct=true&db=s3h&AN=34349143&site=eds-live&scope=site (Publisher: Fitness Information Technology, Inc.)

Paul, R. J., Weinbach, A. P., & Weinbach, C. J. (2003). Fair Bets and Profitability in College Football Gambling. *Journal of Economics & Finance*, *27*(2), 236. Retrieved 2024-01-12, from http://login.proxy01.shawnee.edu/login?url=https://search.ebscohost.com/login.aspx?direct=true&db=bth&AN=9813765&site=eds-live&scope=site (Publisher: Springer Nature) doi: 10.1007/BF02827221

Reed, D. D., Niileksela, C. R., & Kaplan, B. A. (2013). Behavioral Economics. *Behavior Analysis in Practice*, *6*(1), 34–54. Retrieved 2024-02-12, from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3680155/ doi: 10.1007/BF03391790

Schaefer, S. (2011, April). Football betting trends. *Carroll.edu*. Retrieved 2024-04-07, from https://scholars.carroll.edu/handle/20.500.12647/3411

Sinkey, M., & Logan, T. (2014). Does the Hot Hand Drive the Market? Evidence from College Football Betting Markets. *Eastern Economic Journal*, *40*(4), 583–603. Retrieved 2024-03-28, from https://www.jstor.org/stable/24693692 (Publisher: Palgrave Macmillan Journals)

Snaychuk, L. A., Ashley, A. M., McGrath, D. S., Hodgins, D. C., Williams, R. J., & Kim, H. S. (2023, December). Demographic, gambling, and psychological characteristics of sports bettors in a stratified sample of adults who gamble regularly. *Psychology of Addictive Behaviors*, *37*(8), 1089–1100. (Publisher: American Psychological Association) doi: 10.1037/adb0000971.supp

Stetzka, R. M., & Winter, S. (2023). How rational is gambling? *Journal of Economic Surveys*, *37*(4), 1432–1488. Retrieved 2024-02-13, from https://onlinelibrary.wiley.com/doi/abs/10.1111/joes.12473 (_eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/joes.12473) doi:

10.1111/joes.12473

Vandenbruaene, J., Ceuster, M. D., & Annaert, J. (2022, October). Efficient Spread Betting Markets: A Literature Review: Journal of Sports Economics. *Journal of Sports Economics*, *23*(7), 907–949. Retrieved 2024-03-17, from http://login.proxy01.shawnee.edu/login?url=https://search.ebscohost.com/login.aspx?direct=true&db=s3h&AN=158748040&site=eds-live&scope=site (Publisher: Sage Publications Inc.) doi: 10.1177/15270025211071042

Winfree, J. A. (2020, June). Rivalries, Bowl Eligibility, and Scheduling Effects in College Football: Journal of Sports Economics. *Journal of Sports Economics*, *21*(5), 477–492. Retrieved 2024-03-18, from http://login.proxy01.shawnee.edu/login?url=https://search.ebscohost.com/login.aspx?direct=true&db=s3h&AN=142873318&site=eds-live&scope=site (Publisher: Sage Publications Inc.) doi: 10.1177/1527002520916915

Bibliography

Ethan Aaron Reyes

Master of Science of Mathematics

Thesis: GRIDIRON INSIGHTS: PREDICTING GAMEDAY OUTCOMES THROUGH REGRESSION ANALYSIS IN COLLEGE FOOTBALL

Major Field: Mathematics

Education: Bachelor of Science in Mathematics (Baylor University)

Completed the requirements for the Master of Science in Mathematics, Portsmouth, Ohio in July 2024.

7/29/2024

ADVISER'S APPROVAL: Dr. Douglas Darbro

Bibliography

Justin Eugene Beal

Master of Science in Mathematics

Thesis: GRIDIRON INSIGHTS: PREDICTING GAMEDAY OUTCOMES THROUGH REGRESSION ANALYSIS IN COLLEGE FOOTBALL

Major Field: Mathematics

Education: Bachelor of Science in Mathematics (Miami University, Oxford)

Completed the requirements for the Master of Science in Mathematics, Portsmouth, Ohio in July 2024.

7/29/2024

ADVISER'S APPROVAL: Dr. Douglas Darbro