THE PENNSYLVANIA STATE UNIVERSITY
SCHREYER HONORS COLLEGE

DEPARTMENT OF ACTUARIAL SCIENCE

BEATING THE VEGAS SPREAD: USING MULTIPLE REGRESSION TO PREDICT THE OUTCOME OF COLLEGE BASKETBALL GAMES

CONNOR GRABOWSKI
SPRING 2020

A thesis
submitted in partial fulfillment
of the requirements
for a baccalaureate degree
in Actuarial Science
with honors in Actuarial Science

Reviewed and approved* by the following:

Andrew Wiesner
Associate Teaching Professor
Thesis Supervisor

Steven Putterman
Professor of Practice
Honors Adviser

* Electronic approvals are on file.

ABSTRACT

One of the most popular vices in today's society is gambling. Many would argue that no thrill rivals the reveal of a perfect blackjack, when a chosen horse crosses the finish line in first position, or when a roulette spin lands right on what a bettor chose. In recent years, however, there has been an industry-wide shift towards a new style of betting: sports. With legislation passing through the United States Congress early in 2019, sports gambling has become the sexiest avenue for casinos, applications and bookmakers alike. Now with well over $10 billion in total dollar value wagered, gamblers find themselves losing larger and larger sums of money. This study aims to add automation and a mathematical approach to sports gambling through the lens of multiple regression and common programming methods.

In the four major American sports of baseball, basketball, football, and hockey, the three most popular ways to gamble on games are moneyline (predicting the winner), spread (predicting the margin of victory), and over/under (predicting the total number of points, goals, or runs). This study will focus on point spread but explore ways to implement the same mathematical approach to moneyline and over/under. The hypothesis this study aims to test is if a number of ranking metrics that rate teams from first to last can be combined into a single database to predict a single margin of victory for a specific game. This study will focus on the NCAA Men's Basketball tournament from 2014-2019 because of its wide syndication of gamblers as well as college basketball's long list of common rating systems. The study will conclude by testing the database against the results of the 2019 and 2020 NCAA tournaments to determine overall profitability. If this study proves successful, further avenues for implementation in other sports will be explored.

TABLE OF CONTENTS

LIST OF FIGURES iii

LIST OF TABLES iv

ACKNOWLEDGEMENTS v

Chapter 1 Introduction to Sports Gambling 1

Chapter 2 Specific List to Basketball Introduction 4

Chapter 3 Overview of Three Rating Models 7

- 3.1. KenPom 7
- 3.2. RPI 8
- 3.3. BPI 9

Chapter 4 Ties To Actuarial Science 11

- 4.1 Actuarial Science to Gambling Comparison 11
- 4.2. Relation to Society of Actuaries Exams 13

Chapter 5 Data Collection and Model Building 17

- 5.1. Model Example 17
- 5.2. Excel Database 18
- 5.3. R-Studio Model 20

Chapter 6 Commentary on Results 22

Chapter 7 Further Extensions of Study and Conclusions 27

Appendix A RStudio Code 31

BIBLIOGRAPHY 33

LIST OF TABLES

Table 1: Example of Point Spread Bets 6

Table 2: Scenario Analysis 18

Table 3: Selection of Games in Database 23

Table 4: Performance of Selection of Games in Database 24

ACKNOWLEDGEMENTS

I would like to extend my deepest gratitude to Professor Putterman for supporting my ideas and helping me to blend sports and Actuarial Science. I would also like to greatly thank Dr. Wiesner for being such a great resource throughout this whole process and always reminding me that the best data comes from sports.

Chapter 1

Introduction to Sports Gambling

Sports gambling and the United States have a long history together. However, this relationship has only been a legal one for several months (Smiley). Law enforcement and gambling-centric citizens have long had a tug-of-war trying to govern and regulate gambling on horses and sports. Sports gambling bubbled to a head during the 20th century, as organized crime created localized quasi-casinos. To an extent, Congress responded and aimed to rid the nation of what the Los Angeles Times labelled “the worst vice” (Los Angeles Herald). Many years later, the appetite for sports gambling has ballooned, and a growing number of states have given the green-light to sports gambling.

Dating back even further, gambling in the United States existed when the Revolutionary War was funded by government-sponsored lottery drawings. According to the Boston College Law Review, “Two hundred years ago, government-sanctioned lotteries were common throughout America … Lottery proceeds were used to build cities, establish universities, and even to help finance the Revolutionary War” (Rychlak). Furthermore, according to the California Research Bureau, all thirteen original colonies established gambling system that culminated in a $10 million lottery to finance the whole war (Rychlak). This set a precedent with lottery gambling across America that still stands today. Once the nation was officially minted, highly regulated casinos and horse-racing operations sprouted up in many locations. Though these were typically monitored on a strict state-by-state basis, it added a legal vice for many Americans that persisted through hard times like the Civil War, Prohibition, and more.

With regard to sports gambling specifically, due to numerous improprieties of mob-organized gambling and match-fixing scandals like the 1919 Chicago Black Sox, congress tried to keep a tight clamp on the legality, or lack thereof, of sports gambling. The only location where an American citizen could go relieve the urge to gamble on sports was in Las Vegas, Nevada, which was granted the right to administer sports gambling in 1949 in an effort to spur economic activity to the area (Smiley). However, the desire to bet on sports games was widespread across the United States. These gnawings and pangs continued through the 20th and 21st centuries, leading to a monumental Supreme Court decision on May 14, 2018, leaving sports gambling decisions up to the individual states (Smiley). Within six months of this landmark decision by the Supreme Court of the United States, Delaware, Mississippi, New Jersey, Pennsylvania, West Virginia, Rhode Island and Arkansas had fully legalized sports gambling. A flurry of states followed suit, and as of January 2020, twenty states, as well as Washington D.C., were on board. According to the Action Network, a widely syndicated gambling news outlet, 17 more states are expected to pass legislation in the next 24 months (Rovell).

In all of these states, the ways by which a bettor can gamble on a sport or a game are plentiful. For instance, one could bet on the winner, the margin of victory, total points scored, outcomes of future events or championship, individual player stats, and more. It is typically conducted at large casinos at physical sportsbook kiosks, as well as well-known apps like FanDuel and DraftKings. Wagering on sports has been simplified to the touch of a fingertip in many states. The growth of this industry is massive, with over $10 billion in wagered money already (Ramsey). In fact, during the writing of this paper, the 20 billionth legal dollar was expected to be wagered on the upcoming 2020 March Madness tournament prior to the COVID-19 outbreak. By 2025, market experts predict that industry-wide revenue (not wagers) could reach up to $8 billion or $9 billion (Associated Press). States were initially reluctant to legalize another vice, but after completing a simple economic analysis and calculating the potentially lucrative tax returns, they came to appreciate that sports gambling can be a noteworthy addition to the United States economy.

The long and topsy-turvy history of American sports gambling sets the stage for this study as bettors clamor for responsible strategies of sustained profitability. By automating a strategy using statistics and programming, this study aims to aid legal bettors and dispel the notion that "the house will always win." This study will answer the question: can someone consistently gamble and win based on data from a multiple regression model that ranks inputs from first to last?

Chapter 2

Specific List to Basketball Introduction

As previously stated, there is almost no limit to the number of sports, games, or players on which a bettor can place money. From the Major League Baseball (MLB) World Series winner to an National Football League (NFL) running back's rushing yard total to the length of the national anthem prior to the National Basketball Association (NBA) All-Star Game, the 21st-century bettor greatly suffers from the paradox of choice, meaning an overload of options clouds the bettor's vision to make a rational choice. Furthermore, gamblers often suffer from the Monte Carlo fallacy, or the Gambler's Fallacy, which is the mental bias by which bettors anticipate the likelihood of future events based on the frequency of past events (Stafford). This study aims to add some sort of automation to these two and several other mental biases anguishing sports bettors. Do these fallacies actually exist? Is there a way to combine statistics and beat Vegas on a sustained level? Or more simply put, how successful is a gambling regression model solely based on ranking teams from first to last?

At the most basic level, this thesis will test if the idea of automated picks for sports betting is a plausible avenue for long term profitability. By building a computer model that predicts a score outcome along with a predetermined confidence interval, the model will shed some light on whether this style of gambling is any more productive than simply flipping a coin. In the effort of simplicity, the NCAA Men's Basketball Tournament will be the preferred competition of examination. This was selected for a few reasons, with the foremost being that all games are played at a neutral site, allowing the plausibility of this method to be tested without directly having to consider home court advantage. Secondly, college basketball is known for its all-encompassing rating metrics ranking teams from first to last. These help league experts compare teams and predict the annual tournament field, an arduous task for even the most diligent consumer of basketball.

The inputs for this model will be rating statistics that rank teams from first to last. These inputs are ESPN’s Basketball Power Index (BPI), Ken Pomeroy’s KenPom Index, and the NCAA’s Ratings Percentage Index (RPI). All of these statistics will be discussed in greater detail later in the paper, but at their most general level, they use a number of inputs to rank college basketball team efficiency and effectiveness. Most importantly, all three of these models are different ways to assess teams and matchups. By combining these three statistics into one comprehensive regression model, this analysis aims to capture many facets of one game and plot every March Madness game along a complex regression plot and find an accurate point differential.

When speaking about point differential, this is in reference to the point spread, one of the three common methods of betting on basketball. The other two common methods are moneyline (choosing the correct winner) and over/under (how many total points to be scored). When betting on the point spread, the favorite, or team expected to win, is assessed a “minus” value, or a number of points they are expected to win by. For example, if Penn State is favored by the bookmaker to win by three points, Penn State would be considered a -3. In this scenario, the bettor must choose Penn State to win by more than three points or the opposing team to cover, meaning either a win or a loss by fewer than three. Likewise, if Penn State was expected to lose by three, they would be a +3. The bettor then decides to take the favorite to win by their minus value and “cover their spread,” or the underdog to either lose by less than the spread amount or win outright.

For example, let's assume that Duke University has an upcoming game against the University of North Carolina. Vegas casinos give Duke a 3-point advantage, meaning the spread is Duke minus 3. If Duke wins the game by more than 3 points, Duke covers the spread. If Duke wins by 1 or 2 points, UNC covers the spread. Finally, if UNC wins the game, they cover the spread. In the scenario Duke wins by exactly 3 points, this is called a push, meaning the bettor and sportsbook essentially tie. The bettor will then get his or her money returned. This style of betting helps bookmakers to handicap games in an effort to make the odds of winning roughly 50/50. This study aims to find a way to cut through this handicap and earn a long-term profit on betting with point spread, a historically challenging way to play.

Team 1	**Team 2**	**Spread**	**Outcome**	**Cover**
Duke	North Carolina	Duke -3	Duke by 5	Duke
Syracuse	Illinois	Syracuse -10	Syracuse by 3	Illinois
Maryland	Florida	Maryland -1	Florida by 5	Florida

Table 1: Example of Point Spread Bets

It's important to consider that a bookmaker's assessment is not always based on the expected outcome of the game. The bookmakers's primary goal is to create betting opportunities to make money for the sportsbook and hedge against total losses. This is the overall crux of this study. Because of this general ethos of bookmakers, they often adjust spreads, moving them up and down to keep half of the money on both sides of the spread. This kind of betting is not necessarily player performance-based; I hypothesize that this is an inherent market inefficiency that can be taken advantage of.

Chapter 3

Overview of Three Rating Models

As previously stated, to test the plausibility of the idea of whether a multiple regression model based on common college basketball metrics could be successful, this study collected data about college basketball teams for the previous seven seasons. Then a database was created of 5 of these seasons (2014-2018) and tested for success for the two most recent seasons of data. This database was comprised of every first round game in that season's NCAA Men's Basketball Tournament along with the two teams in each game, the score of that game, and the three rankings of each team (RPI, BPI, KenPom). The computer model built in R-Studio created a multiple regression model regressing the predicted point differential (as a plus/minus value) on the six values of ranking (the three rankings of the two teams). In short, we have six seasons of data containing 32 games each that individually contain a score and 6 data points. By inputting the six ranks for an upcoming NCAA tournament game, the model outputs an expected point differential to be compared to a bookmaker's spread. Based on how close or far the predicted spread is from the Vegas spread, bettors could use this model to find optimal betting opportunities. A more complete explanation of the organization of the database will follow later in the write-up, but it's important to note the differences of each individual statistic and how they rate the same outputs with different inputs.

3.1. KenPom

First, KenPom is a basketball index created by professional sports statistician Ken Pomeroy. Data for this oft-cited index is available as far back as 2002 (Pomeroy). In this index, Pomeroy creates several efficiency margins for categories like offense, defense, luck, tempo, strength of schedule, home court advantage, and more. These margins are combined to ultimately create his Adjusted Efficiency Margin, which ends up being KenPom's overall basis of ranking (Pomeroy). Though his actual data collection methods and calculations are proprietary, it's clear that KenPom is a fairly all-encompassing statistic, even factoring in luck and speed of play. In Pomeroy's own words, "The first thing you should know about this system is that it is designed to be purely predictive. If you're looking for a system that rates teams on how "good" their season has been, you've come to the wrong place … The purpose of this system is to show how strong a team would be if it played tonight, independent of injuries or emotional factors. Since nobody can see every team play all (or even most) of their games, this system is designed to give you a snapshot of a team's current level of play" (Pomeroy). One may find this explanation particularly beneficial for college basketball bettors as little to no viewers have the time to watch as many games as his system captures. This system also plays well for March Madness as it predicts how teams would play right now, not ranking who has had the best season thus far. In this model, KenPom was collected for the Selection Sunday (date of tournament selection) for 2014-2019.

3.2. RPI

The NCAA's Rating Percentage Index (RPI) has been used by statisticians and tournament experts, known as bracketologists, since 1984, with its rankings being released weekly since the 2005-2006 season (Sukup). RPI is calculated much differently than KenPom, with more of an emphasis on wins and losses than performance. Again, the specific calculations of the RPI are not available to the public, but the general idea is important to bear in mind. Twenty-five percent of the ranking is based on wins and losses; fifty percent of the calculation is based on strength of schedule (SOS); and the final twenty-five percent is all other opponents' strength of schedule (Sukup). The RPI is essentially a combination of how a team has played, who a team has played, and how tested the opponents are. Whether or not this is a perfect statistic is not what is up for debate in this study. Many would argue that RPI an outdated system, citing that the NCAA prefers to use a new metric known as NET Ranking as of the 2018-2019 season (NCAA.com). However, despite its shortcomings, RPI can be combined with other metrics like KenPom to capture as much of the variation in college basketball games as possible. In the model, RPI was collected for Selection Sunday for 2014-2019.

3.3. BPI

Finally, ESPN's Basketball Power Index (BPI) is the third rating metric to be used in this model. Again, while its official inputs and calculations are about as close-lipped as can be, they give a small explanation about what they are looking for in teams. According to the BPI homepage on ESPN.com, "The College Basketball Power Index (BPI) is a measure of team strength that is meant to be the best predictor of performance going forward. BPI represents how many points above or below average a team is" (Oliver). BPI heavily factors in strength of record (SOR), which is "a measure of team accomplishment based on how difficult a team's W-L record is to achieve" (Oliver). By factoring in both on-court play and strength of schedule, BPI seems to be a hybrid statistic of RPI and KenPom. However, BPI separates itself by factoring in key variables like travel distance, days of rest, altitude, and more. Whereas KenPom and RPI are periodically released rankings, BPI is a daily simulation that updates with 10,000 trials each day. With these final factors like rest and travel considered, BPI attempts to capture any of the remaining variation between two teams in NCAA Tournament games.

Due to BPI being either highly proprietary or simply a limitation of its website, BPI is only available at the end of the season. The model uses end-of-season BPI data. However, to ensure that this data is not significantly skewed by tournament games, a paired T-test will be performed on both KenPom and RPI. By collecting Selection Sunday and end-of-season data for both statistics, this study will ensure that neither statistic significantly changes from dataset to dataset. If both statistics pass this test, we assumed that BPI would produce a similar conclusion.

Chapter 4

Ties To Actuarial Science

While likening actuarial science to sports gambling may seem like a stretch, these two disciplines have much in common. At its core, actuarial science is a field that measures risk, promotes financial security, and analyzes different aspects of loss. Actuaries are often called upon by insurance companies, financial services firms, and other multinational corporations to analyze products, decisions, and other quandaries and assess a risk level to it. How risky is this line of insurance? Can this firm really offer this product for that price? (“Be An Actuary.”) Translating this to the study at hand, the job of an actuary sounds quite a bit like the analysis of choosing games on which to wager money. Does this game seem like a smart gamble? How much risk is tied to this matchup? In the end, sports gamblers and actuaries try to hedge against risk and find the most secure ways to make money. Both groups clamor for reliable tools for automation and risk analysis. That is the overall goal of this project: apply risk characteristics to basketball matchups the same way an actuary may apply risk characteristics to an insurance applicant.

4.1 Actuarial Science to Gambling Comparison

In a recent interview, actuary turned gambling guru Dominic Cortis shared some thoughts about his transition from actuarial risk manager to gambling risk manager. When researching the solvency and risk levels of bookmakers and casinos, he finds one of his biggest struggles, similar to actuaries, is how to “decide a set of assumptions on which you build your model. These make or break the model, so a lot of care needs to be made with respect to choosing, testing and updating these” (Pinnacle). He also finds that individual roles within the two fields often sync up. Some bettors and actuaries focus on model building, while others focus on price inefficiencies, and others focus on data collection and quality assurance. In one of his most interesting arguments, Cortis likens a sports bet directly to insurance, claiming that since a bet is a binary option, it can be thought of as a financial derivative (Pinnacle).

However, there is obviously one key difference between gambling and insuring: insurable interest. When one gambles, he or she pays a premium to have a chance of winning; whereas in insurance, an insured pays a sum of money to indemnify a loss. Cortis explains this with an anecdotal example: “A bet is made to win an amount while an insurance is made to place you in the same position should something occur. That is, if I purchase a contract that pays me a sum on Queen's Elizabeth's death, I am betting. If Prince Philip does it, then he is purchasing insurance” (Pinnacle). As a result of this, bookmakers experience significantly less asymmetric information than in the insurance market. When one purchases insurance, he or she may not fully disclose all information about himself or herself, like some sort of pre-existing cardiovascular condition. On the other side, the insurer could possibly own proprietary information and models that predict lifespan and treatment costs much more precisely than the insured can. This creates a tug-of-war in insurance regarding who has the upper hand and advantage in pricing. In betting, this battle of asymmetric information is much simpler. The bookmaker sets odds with both parties typically having all relevant information available. Barring some sort of insider information on either end, the asymmetric information deficiency is minimized in gambling (Pinnacle). Despite a few key differences like insurable interest and asymmetric information, insurance and gambling have some key similarities that make this study a well-placed exercise within the realm of actuarial science.

4.2. Relation to Society of Actuaries Exams

In addition to this study bridging the gap between Actuarial Science and casino-style gambling, it also can shed some light on several Society of Actuary (SOA) exams. In order to become a fully accredited actuary, candidates must sit for and pass several arduous exams. This process often takes several years, numerous sittings, and hours of study. But, with actuarial science being a field in which correctness is paramount, the SOA's exam-laden process certainly makes a good bit of sense. While some of the preliminary exams seem like more of a math aptitude test than an actuarial problem, this study helps add some real-world perspective to these tests and show how their concepts apply outside of the testing booth.

First, the SOA's exam on Investment and Financial Markets (Exam IFM) tests candidates on "knowledge of the theoretical basis of corporate finance and financial models and the application of those models to insurance and other financial risks" ("Exam IFM: Investment and Financial Markets."). A large portion of this exam revolves around the Efficient Markets Hypothesis (EMH), stating that economies and financial markets are largely efficient, reacting instantly to news and being essentially impossible to beat on a long-term basis. In its most basic form, the EMH has three forms: weak, semi-strong, and strong (Downey). All three forms speak to the quality of financial markets and how they react to information. The weak form states that historic information and data explains how the market will perform. The semi-strong form combines historic information with publicly known information. And finally, the strong form adds private information to the two other sources of information. In the strongest form of the Efficient Markets Hypothesis, the market reacts to previous happenings, information known by the masses, and information known by only a select few (Downey). This study aims to directly apply this idea to gambling, charting how quickly the gambling markets can respond to inputs and if the sports gambling market is beatable in the long run. Can people without private information win in the long run? The EMH would suggest not.

Several papers have been written about this exact idea, the application of Efficient Markets Hypothesis to sports gambling. One such study completed by David N. DeJong of the University of Pittsburgh titled “Using Past Performance to Predict NFL Outcomes: A Chartist Approach” used a variety of betting strategies to chart outcomes in NFL games in the late 1990s (DeJong). He found that his results reflect poorly on the Efficient Markets Hypothesis. He concluded that his conservative strategy “generated 65 wins in 103 bets over this period -- a 63.1-percent success rate -- which leads to a rejection of the break-even hypothesis at the 3-percent significance level” (DeJong). Furthermore, another one of his gambling strategies hit at a 58-percent rate, well above the percentage needed for long-run success. That being said, DeJong makes an interesting conclusion about the publishing of his paper. By spreading this information and his processes, he is simply adding to public information that will now be captured by the betting market and corrected by the EMH (DeJong). While his strategies worked in the interim, can they be applied over multiple seasons? Or has the publishing of this paper directly corrected for the exact inefficiency he was commenting on? There’s reason to believe that if one feels that he or she has found a way to make long-term profit, he or she should keep it close-lipped. By making this information public, bookmakers would likely be able to make adjustments. In this regard, successful private information would violate the strong form of the EMH for the lucky few.

Fast forwarding nearly a decade and switching sports, Rodney Paul and Andrew Weinbach wrote a paper in 2005 for the Journal of Economics and Finance titled “Market Efficiency and NCAA College Basketball Gambling” (Paul). In this paper, they test a similar theory to DeJong and employ a black-and-white strategy to certain betting situations over several seasons. According to their findings, “The betting market for NCAA college basketball is examined from the 1996-97 season through 2003-04. In the overall sample, market efficiency cannot be rejected. For big favorites, specifically those favorites of 20 or more, a simple strategy of betting the underdog in these games is shown to reject the null hypothesis of a fair bet since the underdog wins more than implied by efficiency” (Paul). They concluded that bettors have an inherent bias to bet on favorites and home teams that allow bookmakers to add points to spreads and collect exorbitant profits on these biases. This allows for the EMH to hold at the weak and semi-strong forms, at least, due to historical information and public information swaying bettors to make decisions and ultimately lose more than they win. This study does not aim to seek out betting biases and capitalize like Paul and Weinbach, but to simply develop a computer regression system that predicts game outcomes based on numerous games consisting of teams with similar ratings and statistical qualities. By and large, this seems to be a methodology and system not yet largely tested.

In addition to its almost direct application to Exam IFM, this study has qualities that can help explain other SOA exams. In the Statistics for Risk Modeling Exam (Exam SRM), students are asked to “be familiar with regression models (including the generalized linear model), time series models, principal components analysis, decision trees, and cluster analysis. Candidates will also be able to apply methods for selecting and validating models” (“Exam SRM: Statistics for Risk Modeling.”). This study is an extensive exercise in multiple regression modeling as well as model selection and validation. In addition to developing a college basketball model, the study aims to perfect, tweak and assess the findings, validating its quality and potentially applying it to other similar sports. In the SOA’s Predictive Analytics Exam (Exam PA), candidates are tasked with a long business problem to solve, including solving for data manipulation and R-studio regression ("Exam PA: Predictive Analytics."). This study also includes both of these tasks, gathering a multitude of data from a number of sources, scrubbing and cleaning it, and gleaning meaningful results within R-studio. In the end, this college basketball regression test does not seem a problem with directly actuarial roots, but by comparing actuarial science to gambling and likening parts of the process to SOA exams, one will quickly see the breadth of actuarial thinking, and how far actuarial skills can span in the real world.

Chapter 5

Data Collection and Model Building

To quickly recap how the model and betting tool will work, users will find a March Madness game he or she would like to bet. The bettor will then look up the three key statistics of KenPom, RPI and BPI for the two teams on each of the statistics websites, which are updated either daily or weekly. The bettor then inputs the six statistics into a multiple regression model that regresses on numerous March Madness games from the past five to six seasons. The model will analyze the inputs, place them somewhere within the extensive database, and output an estimated point differential along with a 95% confidence interval. The point differential will dictate which side of the spread to bet on, and the confidence interval will provide goalposts for optimal betting opportunities. Finally, the bettor will analyze his or her Vegas spread and decide which way to bet, or if to even bet at all.

5.1. Model Example

This process is a bit complex, so it is best to illustrate with an example. Disclaimer, these numbers and the scenario are purely hypothetical. This previous March, a gambler wanted to potentially place money on a game between Penn State and Pittsburgh. The bettor looked up Penn State's statistics of (30 KenPom, 26 RPI, and 34 BPI) along with Pittsburgh's statistics of (71 KenPom, 59 RPI, 82 BPI). The bettor then adds these inputs to the regression model. The program output a point differential of 5.738 and a 95% confidence interval of about 2.7 to 8.7. This output means the program anticipates Team 1 (Penn State) will win by 5.738 points, based on the log of games in the database. The program is also 95% confident that Penn State should be favored by between 2.7 to 8.7. If Vegas were to offer a spread higher than 8.7 points or lower than 2.7 points, the program would consider this to be an optimal betting opportunity. Though some games may not be considered optimal bets, they should not be fully dismissed. In fact, the profitability of a variety of strategies will be analyzed later in the writeup. The below chart illustrates different scenarios with how this game could play out given the model's output of PSU -5.74 and (PSU -2.7, PSU -8.7).

Vegas Spread	Model Spread	Model CI	Model's Choice	Optimal Bet?	Outcome	Model Win or Loss
PSU -11	-5.74	(-2.7 to -8.7)	Pitt +11	Yes	PSU by 5	Win
PSU -6	-5.74	(-2.7 to -8.7)	Pitt +6	No	PSU by 5	Win
PSU -5.5	-5.74	(-2.7 to -8.7)	PSU -5.5	No	PSU by 5	Loss
PSU -2	-5.74	(-2.7 to -8.7)	PSU -2	Yes	PSU by 5	Win

Table 2: Scenario Analysis

5.2. Excel Database

Before testing the model and database, it is important to understand how both are comprised and organized. The database itself is housed in Microsoft Excel before converted to a .CSV file to become readable in R. The Excel file has four sheets for each year with data collection: 2019 through 2014. The 2020 NCAA Tournament was expected to be included in the database, but the event was canceled due to the COVID-19 outbreak. Though this may seem somewhat detrimental to the study at hand, there remains enough data to proceed and draw reasonable conclusions. The model test will run the 2019 tournament and a variety of regular season games that happened in early 2020. While these 2020 games were not initially intended to be included in the study as they do not take place within the NCAA Tournament, they will be included to reinforce and/or bolster the results captured in the 2019 Tournament.

As previously stated, each year of data has four spreadsheets: a rollup of data, an RPI sheet, a KenPom sheet, and a BPI sheet. Each of the three statistical sheets was carefully transferred from their individual sites, either downloaded as a PDF or copy-pasted. For example, for the 2016 BPI sheet, the data was downloaded from ESPN’s BPI archive and copied to the proper sheet. In the same way, the 2019 RPI sheet was created by converting the NCAA’s PDF of data into a readable Excel format. This general process was followed for the three statistics for the six seasons. However, a small extra step was taken for KenPom and RPI. Discussed earlier, BPI is only available as an end-of-season statistic. To be able to include this value in the model, it must be concluded that BPI does not significantly differ from the tournament’s selection day to the end of the season. To test this, KenPom data and RPI data will each undergo a paired t-test to check this assumption. Being that BPI has similar inputs as KenPom and RPI, if they pass the test, one could reasonably assume that BPI would as well. In preparation for this test in R-Studio, end of season data was also accumulated for all KenPom and RPI sheets.

After making the individual data sheets and matching selection data to end of season data, a rollup sheet for each season’s tournament was created to include the 32 first round matchups, the scores, point differential (what is being predicted by the model), and the six statistics copied in via lookup. Finally, each of the six rollup sheets combines to form one final master sheet that is read into R-Studio. This final master sheet is the regression model to be tested. This sheet has 6 years of 32 games as well as the following data columns: Year, Team1, Team2, Team1Points, Team2Points, PointDiff, RPI1Final, RPI2Final, KenPom1Final, KenPom2Final, BPI1, BPI2, RPI1Selection, RPI2Selection, KenPom1Selection, and KenPom2Selection.

5.3. R-Studio Model

In order to assume the BPI end of season numbers did not significantly change from prior to the start of the tournament, we conducted a paired test on KenPom and RPI. We used the pair t-test to test for significant changes in index rankings of teams prior to start of the tournament to their ranking at tournament conclusion. A paired t-test, which is a statistical test checking how significant a group of variables differ from trial 1 to trial 2, will be used in this test. We are looking for an insignificant result at the 0.05 significance level, meaning that from collection 1 to collection 2, the data stays relatively the same. For example, say a group of people's weight is tested before and after a diet. If the paired t-test has a p-value of 0.05 or lower, we can conclude that there is a significant difference in subjects' weights from start of diet to end of diet. If the paired t-test has a p-value of higher than 0.05, we cannot conclude that there is a significant difference in subjects' weights. When running this same test for five seasons of KenPom data and excluding 2019 as that is the year to be tested, we attained a p-value of 0.1536. This means we cannot conclude that there is a significant difference in KenPom data from day of selection to end of season. When running the test for RPI, we attained a far more insignificant result, with a p-value of 0.6236. Since both statistics pass the test for insignificance, we can apply the same assumption for BPI and proceed with all three variables in the study.

After completing both assumption tests, we proceed with the model creation. While the code for this exercise is moderately simple, it is included in Appendix A. After reading in the large master sheet, variable names are assigned for each key value and the regression model is built. There are the slots for the six key inputs mentioned throughout the writeup, as well as a seventh input to toggle various levels of confidence. The default confidence level to determine optimal bets is 95%, but this can be increased or decreased to a bettor's liking and level of risk aversion. The model has a multiple R-squared value of 0.4332, meaning that over 43% of the variation in point differential is captured by the model. The model also has an overall p-value of less than 0.001 on 153 degrees of freedom, meaning the model is a significant predictor of point differential as the p-value is far lower than 0.05. From a statistical perspective, the model appears to hold up in terms of robustness, but the best indicator is how it performs against real basketball games.

Chapter 6

Commentary on Results

To properly delve into the performance of the model, one must take a three-pronged approach to the analyses. First, where is the team and gambling data sourced from? Much of this has already been covered, but it is important to know how Vegas spreads are created and modified. Secondly, how many wins and losses did the model achieve? And finally, what level of profitability does this lead to? This step is far more important than wins and losses, because breaking even with wins and losses in the long term could lead to significant monetary losses from the bettor's perspective. In addition to looking at the spread, one must also look at the Vegas odds. Typically, it is set at -110, meaning you need to bet $11 to win $10, or bet $110 to win $100. This is sometimes called "juice" by gamblers, and is how the casinos are able to tack on a bit more profit and loss indemnity to their books. In this scenario, one must make sure that, with a consistent betting unit, the model not only wins more often than loses, but also makes money.

Specific odds and spread data for most games are not readily available for many casinos and bookmakers, especially since sports gambling is a relatively new source of industry. However, some sources track various lines, spreads, and odds in virtual databases from back when gambling was more of a black-market vice. Many gamblers like to peruse how certain teams perform against the spread, sometimes referred to as ATS. This is basically just a compilation of how many times a team covered the spread versus how many times the team failed to cover the spread. A team's performance ATS can be analyzed over a specified period of time or within a set of parameters, like home games or games that start during the daytime. Furthermore, there are often significant trends like home teams covering the spread when favored, or road underdogs covering the spread when given more than 10 points. To track these trends and many more points of inquiry, the Odds Shark NCAAB Basketball Odds & Handicapping Database is a great place to start. Recalling back to earlier in the paper, this study is a plausibility check more than anything else, meaning that a perfect tracking of lines and spreads is not necessarily vital. Odds Shark has the closing spread for the 2018 tournament games. While a bettor may be able to find better success with other casinos, this was a strong and effective starting point to verify the plausibility of this gambling style ("NCAAB Basketball Odds & Handicapping Database."). Below is a sample of how some of the games were charted for the study, and how they played out:

Team 1	Team 2	Spread	Outcome	Winner	Team
Duke	ND State	-27.5	-23	Underdog	ND State
VCU	UCF	-1.5	+15	Underdog	UCF
Miss St.	Liberty	-7	+4	Underdog	Liberty
Virginia Tech	Saint Louis	-10	-14	Favorite	Virginia Tech
Maryland	Belmont	-3.5	-2	Underdog	Belmont
LSU	Yale	-7	-5	Underdog	Yale
Louisville	Minnesota	-5.5	+10	Underdog	Minnesota
Michigan St	Bradley	-17.5	-11	Underdog	Bradley
Gonzaga	FDU	-28.5	-38	Favorite	Gonzaga

Table 3: Selection of Games in Database

As a quick example, Duke played North Dakota State. Duke was favored to win by 27.5 points, according to Odds Shark. The outcome of the game was that Duke won by 23 points. Since 23 points is less than 27.5 points, North Dakota State (the underdog) covered the spread despite Duke winning the game. After recording the game data for these 9 games and the 23 other first round games analyzed, it was then time to input this into the model and see how well a bettor would have been able to predict these outcomes. Also while 32 games seem like a remarkably small sample size, the initial intent of this experiment was to have a 64 game test before the COVID-19 outbreak. Returning to the previous example, Duke's statistics, in the order of KenPom, RPI, BPI, went (3, 3, 3). North Dakota State's statistics were (199, 222, 193) For this game, the model predicted that Duke should defeat North Dakota State by 15.3 points, with a 95% confidence interval of 4.8 points to 25.8 points. Since Odds Shark's line of 27.5 points falls outside of the confidence interval of (4.8 to 25.8), this was considered an optimal betting opportunity. The model is 95% confident that an outcome of 27.5 points would not fall within the interval of possibility. The model's prediction was ultimately correct, since it calculated the outcome on the winning side of the spread. In short, the model predicted this game would be closer than the spread was offering, thus one should bet on North Dakota State. The same games as above are shown below with their model predictions and ultimate outcomes:

Model Pick	**Model's Team**	**95% Confidence**	**Game Outcome**	**Win or Loss**	**Optimal?**
Duke -15.3	ND State	-4.8 to -25.8	-23	W	*
UCF -3.9	UCF	-8.4 to +0.6	+15	W	*
Miss St -4.1	Liberty	-2.1 to -6.1	+4	W	*
VT - 11	VT	-8.4 to -13.5	-14	W	
MD -1.4	Belmont	-3.6 to +.7	-2	W	
LSU -6.3	Yale	-4.1 to -8.4	-5	W	
Louisville -3.8	Minnesota	-1.2 to -6.45	+10	W	
MSU -22.7	MSU	-18 to -27.5	-11	L	
Gonzaga -25.1	FDU	19.73 to -30.5	-38	L	

Table 4: Performance of Selection of Games in Database

Seen above, the model performed quite well with the above games, winning seven out of the first nine matchups. Furthermore, the three games that the model marked as optimal betting opportunities all ended up as wins. This trend of success carried throughout the 32 first round games of the 2019 NCAA Men's Basketball Tournament, with the model winning 22 of the 32 games. Even more impressively, the model predicted seven optimal betting opportunities, winning all seven of them. In the end, the model's 22 wins equates to a 69% win percentage, with a perfect 100% success rate on optimal games.

To bolster these results, the study aimed to also test the 2020 NCAA Men's Basketball Tournament, but with the tournament being canceled, we must search for other sources of data. Once intended to be a fun test and not included in the write-up, a variety of regular season games in late February and early March were thrown into the model after it was coded in R-Studio. These games also had the added benefit of 2019 data included in the database, something that was obviously not in play for the previous season. The model performed similarly, winning 19 out of 30 games, roughly 63%. Again, while this second set of results should not be taken at face value, it helps back up the performance of the 2019 Tournament test - a hit rate that one should certainly have a great deal of skepticism over.

While a win rate of almost 70% sounds great, how does that translate to profits? This depends on the strategy a bettor wants to take. Three key strategies jump out at first glance: betting one unit on every tested game, betting one unit on every optimal game, and dividing 32 total units evenly over all optimal games. With a set base unit of $100 per game, the three strategies would be one of $100 each on 32 games, $100 each on 7 games, or about $450 each ($100 x 32 / 7) on 7 games. These three strategies are met with different risk levels, different total winnings, and different returns on investment. For simplicity's sake, it is also to be assumed that each game has odds of -110. While games often post odds of -105 (bet 105 to win 100) or -100 (bet 100 to win 100), this conservative approach will provide for the safest estimate of profits. The first strategy, a $100 wager (or a bet of $110) on each game provides a total profit of $1100 and a return on investment (ROI) of 31.25%. The second strategy, a $100 wager on the seven optimal games provides a total profit of a perfect $700 and a profit margin of 90.91%. Finally, the more lucrative version of the second strategy, a wager of about $450 on each of the 7 optimal games, returns a staggering profit of $3,150, and the same profit margin of 90.91%.

While these results, profits, and ROIs are not guaranteed to be replicated every test, tournament, or season, they certainly prove the overall plausibility of this style of college basketball gambling. Though sample size remains small, it seems very much possible that a bettor could gather various rating metrics about two teams, create a database of past games, and predict gambling outcomes with varying levels of certainty. Many argue that gambling should be whittled down to a coin flip, as breaking even is seen as the long-term outcome. However, this multiple regression database shows early signs of bucking that trend, winning significantly more than losing, and finding a market inefficiency that should be capitalized on going forward.

Chapter 7

Further Extensions of Study and Conclusions

As determined and calculated in Chapter 6, this style of gambling can surely have success in the immediate short term. While unforeseen circumstances prevented a longer term analysis, the early results on this train of thought are undoubtedly positive. Bearing this in mind, there are several modifications or tweaks that could be explored to improve the quality of the database and its results. First and foremost, something that could have certainly improved the accuracy of the regression model would be to use actual index values instead of index rankings. Something like KenPom is obviously ranked, but behind that first-to-last ranking is calculated number based on various inputs. These calculated numbers are then ordered for ranking. For example, let's take a game that was earlier analyzed like Duke vs. North Dakota State. Instead of using Duke's statistics of (3, 3, 3), one could have better calculated the spread if instead of rankings, the database used hard numbers with precise decimals. However, this violates the initial intent of this study, testing how accurate a gambling model could be that simply ranks inputs from first to last. Despite this being outside of the study's scope, it would have made for more mathematically sound results. Perhaps some other adjustment could have been made for teams coming into the tournament on a significant winning or losing streak, despite that likely accounted for in the three rating metrics. A deeper study into tournament patterns could also examine how teams from certain conferences perform, similar to how an actuary or underwriter examines demographic information for prospective insureds.

Another way to reduce error in the study would be to add more rating metrics to the database. Not only does three statistics feel like a "magic number" to have believable and robust results, it was also the extent of widely syndicated statistics that were easily accessible. In the future, I would like to examine the Sagarin metric and the NCAA's Net Rating, but neither of these were attainable for the study at hand. Sagarin's data is not archived online, and Net Rating has only been used in practice for two seasons. Of course, there was the thought of developing a model-specific statistic, but with the intent of the study being to develop a statistical model, not an extremely complex statistic itself, this remained in the "Future Considerations" section for now. This line of thought mirrors how actuaries and insurance professionals constantly clamor for more ways to examine risk, assess policies, and angle for new conclusions.

With the database running a multiple mean squared regression model, there was always the thought of manipulating the style of regression, or how the model reads in variables. Since basketball scores and ranking statistics are all integers, perhaps a Poisson style of regression could work. However, since Poisson regression predicts the individual counts of an event happening, there is no room for negative values. There would need to be some sort of log-transformation or other manipulation. Perhaps after running the model for hundreds of games, patterns could emerge, and game scores could end up following some sort of normal distribution. If that ended up being the case, we could take more of a probabilistic approach, incorporating z-scores and percentiles.

Finally, what about how the regression model interprets the interaction of the variables? Right now, the model takes the six individual numbers and tries to place this game somewhere in the plethora of games in the database. However, the distribution of scores could provide to be a bit more uniform if instead of looking at (A, B, C) vs. (D, E, F), we took the approach of (A-D, B-E, C-F). Meaning, instead of inputting both ratings, we input the difference between the two ratings. Then, this six-piece regression model would become a three-piece model. It would be an interesting point of study to discover if something like 1 vs. 51 is much different than 101 vs.

151. If the difference is negligible, then the three-piece model could work much better. When examining the results, one can see that games that have large differences in ratings lead to wide error margins, for example the Duke vs. North Dakota State game. By simplifying the model, more games could be treated similarly, potentially reducing error and creating a more condensed database without having to add any games. Taking this once again to an insurance perspective, insurers prefer to have as few risk buckets as possible, to keep calculations and pricing simple.

An analysis like this could very well be applied to other sports as well. Though college basketball provides for the strongest test case because of the pomp and circumstance of the NCAA Tournament as well as its expansive swath of teams, other sports could similarly follow suit. For example, if one charted 6 or 7 years of NFL or College Football scores along with the rank of both teams' offensive efficiency, defensive efficiency, and special teams efficiency, a similar football gambling database would form. Due to the exorbitant number of games in a season, baseball would be a tougher analysis to perform; however, the process remains the same. One would need to collect data on every game for 5-7 seasons, along with each team's rank in offense, defense, and pitching, and then complete the previous database analysis. Forgetting about sports altogether, what if insurers completed this style of analysis? Let's assume Insurer XYZ has 10,000 insureds and decided to rank the carefully calculated categories of claim risk, financial risk, longevity risk, variability risk, etc. from 1 to 10,000. They could create a database of all these calculated categories along with their annual gross claim total paid. Then using this database to assess new clients, insureds can have another way to calculate expected value of future claims. They would simply input the new client's risk levels and would receive an expected claim amount. Though this is likely overly simplified for large insurers with complex pricing tools, the framework of the analysis could certainly be implemented in some way, shape, or form.

Back in Chapter 1, the question of "Can someone consistently gamble and win based on data from a multiple regression model that ranks inputs from first to last?" was posed. Though sample size remains small, one could surmise with a moderate level of caution that the answer is yes. With ROI margins ranging from 30% to 90%, this certainly seems like a tenable strategy, at least in the short term. In proving the model's success, one's mind can race about other modifications, tweaks, and changes that could turn this model into a living and breathing picks machine. Perhaps it could work for other sports, regular season games, or even the insurance industry as a whole. Despite a great deal of interference from the COVID-19 outbreak, there certainly remains a great deal of optimism about the strategy and future expansions of this study.

Appendix A

RStudio Code

#Setting Working Directory
```
setwd("Documents")
setwd("Thesis")
```

#RPI Assumption Test
```
RPITestR = read.table('rpitestr.csv', sep=',', header=T)
print(RPITestR)
RPI1 <- RPITestR[[1]]
RPI2 <- RPITestR[[2]]
t.test(RPI1, RPI2, paired=TRUE)
```

#KenPom Assumption Test
```
KenPomTestR = read.table('kenpomtestr.csv', sep=',', header=T)
print(KenPomTestR)
KenPom1 <- KenPomTestR[[1]]
KenPom2 <- KenPomTestR[[2]]
t.test(KenPom1, KenPom2, paired=TRUE)
```

#Reading In R Database and Assigning Variables
```
RDatabase = read.table('rdatabasetest2019.csv', sep=',', header=T)
print(RDatabase)
Year <- RDatabase[[1]]
Team1 <- RDatabase[[2]]
Team2 <- RDatabase[[3]]
Team1Points <- RDatabase[[4]]
Team2Points <- RDatabase[[5]]
PointDiff <- RDatabase[[6]]
RPI1Final <- RDatabase[[7]]
RPI2Final <- RDatabase[[8]]
KenPom1Final <- RDatabase[[9]]
KenPom2Final <- RDatabase[[10]]
BPI1 <- RDatabase[[11]]
BPI2 <- RDatabase[[12]]
RPI1Selection <- RDatabase[[13]]
RPI2Selection <- RDatabase[[14]]
KenPom1Selection <- RDatabase[[15]]
KenPom2Selection <- RDatabase[[16]]
```

#Creating Model
```
library(car)
model = lm(PointDiff ~ RPI1Selection + RPI2Selection
```

```
                    + KenPom1Selection + KenPom2Selection
                    + BPI1 + BPI2, data=RDatabase)
summary(model)
predict(model, interval="confidence", level=.95, se.fit=T,
       newdata=data.frame(RPI1Selection = 4,
                    RPI2Selection= 157,
                    KenPom1Selection=8,
                    KenPom2Selection=152,
                    BPI1=7,
                    BPI2=150))
```

BIBLIOGRAPHY

Associated Press. "Sports Betting Market Expected to Reach $8 Billion by 2025." *MarketWatch*, MarketWatch, 4 Nov. 2019, www.marketwatch.com/story/firms-say-sports-betting-market-to-reach-8-billion-by-2025-2019-11-04.

"Be an Actuary." *What Is an Actuary?*, Society of Actuaries, www.beanactuary.org/what/.

DeJong, David N. "Using Past Performance to Predict NFL Outcomes: A Chartist Approach." *Semantics Scholar*, University of Pittsburgh, Mar. 1997, pdfs.semanticscholar.org/1c08/91e5a066ba4b6a56f82d4e90397bd6842091.pdf.

Downey, Lucas. "Efficient Market Hypothesis (EMH)." *Investopedia*, Dotdash, 2 Feb. 2020, www.investopedia.com/terms/e/efficientmarkethypothesis.asp.

"Exam IFM: Investment and Financial Markets." *SOA*, Society of Actuaries, www.soa.org/education/exam-req/edu-exam-ifm-detail/.

"Exam PA: Predictive Analytics." *SOA*, Society of Actuaries, www.soa.org/education/exam-req/edu-exam-pa-detail/.

"Exam SRM: Statistics for Risk Modeling." *SOA*, Society of Actuaries, www.soa.org/education/exam-req/edu-exam-srm-detail/.

"Los Angeles Herald, Volume XXXIX, Number 181, 30 April 1913." *Los Angeles Herald 30 April 1913 - California Digital Newspaper Collection*, DL Consulting, cdnc.ucr.edu/?a=d&d=LAH19130430.2.57&e=-------en--20--1--txt-txIN--------1

NCAA.com. "The NET, Explained: NCAA Adopts New College Basketball Ranking." *NCAA.com*, Turner Sports Interactive, 22 Aug. 2018, www.ncaa.com/news/basketball-men/article/2018-11-26/net-explained-ncaa-adopts-new-college-basketball-ranking.

"NCAAB Basketball Odds & Handicapping Database." *Odds Shark*, Odds Shark, www.oddsshark.com/ncaab/database.

Oliver, Dean. "Introducing the BPI." *ESPN*, ESPN Internet Ventures, 11 Feb. 2012, www.espn.com/mens-college-basketball/story/_/id/7561413/bpi-college-basketball-power-index-explained.

Paul, Rodney, and Andrew Weinbach. "Market Efficiency and NCAA College Basketball Gambling." *Journal of Economics and Finance*, Springer US, Sept. 2005, link.springer.com/article/10.1007/BF02761584.

Pinnacle. "From Actuary to Sports Betting." *Pinnacle*, Ragnarok Corporation N.V., 20 Jan. 2016, www.pinnacle.com/en/betting-articles/educational/from-actuary-to-sports-betting/VLQ2VSYXLBX62L89.

Pomeroy, Ken. "Ratings Explanation." *Kenpom.com*, The Forecast Factory LLC, 29 Nov. 2006, kenpom.com/blog/ratings-explanation/.

Ramsey, Eric. "The First $10 Billion In Expanded US Sports Betting Revenue Goes To ..." *Legal Sports Report*, 27 Nov. 2019, www.legalsportsreport.com/35373/winners-losers-10-billion-us-sports-betting/.

Rovell, Darren. "Legalization Tracker: Washington Becomes 21st State to Legalize Sports Betting." *Action Network*, The Action Network, 26 Mar. 2020, www.actionnetwork.com/news/legal-sports-betting-united-states-projections.

Rychlak, Ronald. “Lotteries, Revenues and Social Costs: A Historical Examination of State-Sponsored Gambling.” *Boston College Law Review*, Boston College, 1992, lawdigitalcommons.bc.edu/cgi/viewcontent.cgi?article=1964&context=bclr.

Smiley, Brett. “A History of Sports Betting in the United States: Gambling Laws and Outlaws.” *SportsHandle*, 13 Nov. 2007, sportshandle.com/gambling-laws-legislation-united-states-history/.

Stafford, Tom. “Why We Gamble like Monkeys.” *BBC Future*, BBC, 27 Jan. 2015, www.bbc.com/future/article/20150127-why-we-gamble-like-monkeys?referer=https%3A%2F%2Fen.wikipedia.org%2F.

Sukup, Jim. “What Is the RPI?” *RPI Reports*, Collegiate Basketball News Company, rpiratings.com/WhatisRPI.php.

ACADEMIC VITA

CONNOR T. GRABOWSKI

ctg5142@psu.edu

EDUCATION

The Pennsylvania State University | Schreyer Honors College | Class of 2020 | University Park, PA
Smeal College of Business *B.S. in Actuarial Science* | Eberly College of Science *Minor in Statistics*

ACTUARIAL EXAM PROGRESS

Exam 1/P – Probability: Pass: January 2018
Exam 2/FM – Financial Mathematics: Pass: March 2019
Exam 3/IFM – Investment and Financial Markets Pass: November 2019
Economics, Accounting and Finance VEE: Completed: December 2017

WORK EXPERIENCE

Mercer **Philadelphia, PA**
Actuarial Intern, Health & Benefits *May 2019 – Aug 2019*
- Assisted various health and benefit consultant teams servicing self-insuring businesses in Greater Philadelphia
- Compiled monthly claim and enrollment data metrics to create visuals and presentations delivered to clients
- Created 2020 plan documents for large client determining various contribution levels per plan option and facility

AXA-XL **Exton, PA**
Actuarial Intern, North America Pricing *May 2018 – Nov 2018*
- Assisted Design Professional team to correct negative annual rate change in their small account pricing tool
- Prepared 2019 Base Loss Ratio Plan for executive meeting with senior leaders of Pricing and Reserving teams
- Analyzed monthly rate monitor data for Environmental team to improve quality and efficiency of Access database

Duff & Phelps **New York, NY**
Intern, Marketing Department *May 2017 – Nov 2017*
- Assisted in implementation of new website by working closely with back-end developers and digital team
- Completed regular checks of third-party SEO programs and examined web traffic data to optimize live webpages
- Drafted and published over 50 web pages for news and publications on Percussion content management system

Trenton Thunder **Trenton, NJ**
Gameday Statistics Specialist *Feb 2015 – Aug 2017*
- Collaborated with MLB and directed online user interface to provide instantaneous live statistics from games
- Coordinated with Director of Creative Services in statistics department for NY Yankees Minor League affiliate
- Managed Daktronics Scoreboard and ensured correct electronic translation of numeric values during games

---

LEADERSHIP EXPERIENCE

Actuarial Science R Programming Club **University Park, PA**

Founder *July 2018 – Present*

- Designed curriculum using DataCamp to self-teach cohort of Penn State Actuarial Science students RStudio
- Organized weekly two-hour meetings to cover topics like manipulation, visualization and TidyVerse packages
- Developed intermediary proficiency in R geared towards being applicable to relevant actuarial work

---

AWARDS AND RECOGNITION

- Smeal College of Business Sam Wherry Honors Scholarship (2016, 2017, 2018)
- Schreyer Honors College Academic Excellence Scholar (2016, 2017, 2018), Study Abroad Travel Grant (2018)
- TEAMS (Tests in Engineering, Aptitude, Math, Science), National Finalist (2013 & 2016)
- Interests: Sports Analytics (esp. English Premier League Soccer & NBA), Travel (Semester Abroad Spring 2018)
- Delta Sigma Pi Professional Fraternity Chair of Social Activities and Epsilon Plendge Class Vice President (2017)