THE PENNSYLVANIA STATE UNIVERSITY
SCHREYER HONORS COLLEGE

DEPARTMENT OF STATISTICS

PREDICTING THE OUTCOME OF THE FINAL SIXTEEN TEAMS IN COLLEGE BASKETBALL USING TIME SERIES ANALYSIS AND MARKOV CHAINS

CHRISTOPER KENNETH MILLER
SPRING 2023

A thesis
submitted in partial fulfillment
of the requirements
for baccalaureate degrees
in Statistics and Mathematics
with honors in Statistics

Reviewed and approved* by the following:

Andrew Wiesner
Associate Teaching Professor of Statistics
Thesis Supervisor

David R. Hunter
Professor of Statistics
Honors Adviser

* Electronic approvals are on file.

ABSTRACT

Dummy Abstract Dummy Abstract.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

Acknowledgments Acknowledgments Acknowledgments Acknowledgments Acknowledgments Acknowledgments Acknowledgments Acknowledgments Acknowledgments Acknowledgments

Acknowledgments Acknowledgments Acknowledgments Acknowledgments Acknowledgments

Acknowledgments Acknowledgments Acknowledgments Acknowledgments Acknowledgments

Chapter 1

Introduction

National Collegiate Athletic Association (NCAA) Division I men's basketball is a widely publicized sport in the United States, and like many other sports, the use of statistics to make some informed decision has become relevant for the league, analysts, coaches, and even the casual fan. The validity and usefulness of certain statistically backed reasoning can be questioned and explored but it is undeniable that number-backed decisions provide concreteness to any given conclusion. In the eyes of analysis, this sport provides an extra level of difficulty due to the collegiate aspect. In a traditional professional league, you can note some small material differences between teams, for example, payroll size, organization location, owner investment strategies, but overall, you can do analysis with the assumption that the professional teams are all on the same level of fairness. When we look at the collegiate level there are two main levels of unfairness that we must consider when looking to do any sort of statistical analysis, access to funds and recruiting level.

It is known that the NCAA basketball tournament, also known as March Madness , the final tournament of the NCAA season is regarded as the pinnacle of sports. It is a single elimination tournament that decides the overall champion of college basketball each year. The tournament is set up such that out of the 32 Division I conferences, the champion of each is guaranteed a spot in the tournament, then 36 other teams that impressed the NCAA committee(1). This guarantee's representation of every conference, and rewards teams that play tougher conferences, which goes back to the unfairness factor in this sport. When the field of 68 teams is set, the NCAA committee then decides seeding, such that the best teams would play the worst teams on a path to the championship, this rewards the teams that did the best in the regular season. This seeding decision by the NCAA is at least in part, statistically based, and by creating an order of teams, the NCAA is essentially making their own prediction of what teams they think are better than others(1). If the NCAA's ranking was completely true, then the lower seed would always win with the top ranked number one seed winning the whole tournament. We know this is not true, for example, since 1984 when the tournament expanded to 68 teams the seed 5 teams only have a 63% win rate in the initial matchup against the seed 12 teams(2). Much of this randomness in predicting relative team performance has to do with the complexity of valuing how much the aforementioned unfairness contributed to the team's performance.

Sample

Motivation and Overview

Sample

Chapter 2

Efficiency

Sample

Chapter 3

Data Considerations and Setup

Sample

Chapter 4

Time-Series Model

Sample

Chapter 5

Markov Chain as Probability Model

For every year we have a set of numbers defining the teams to participate in the final NCAA tournament. As stated in the previous section it was stated how we are using a time-series analysis to populate these set of numbers, but we can see that the following markov chain method to get expected probabilities of teams making it to the second and third rounds can be applied to any set of numbers that are defining the teams as long as those set of numbers are positive. This idea of a probabilistic model to create expected results will loosely follow the work of (SCHWERTMAN)

There are 64 teams in the tournament and these teams are split into four different regions. For any given year we are predicting the last four remaining teams in each region therefore overall predicting 16 teams. This stochastic method takes the set of any pair of four connected teams in any given region, for example, seeds (1,16,8,9) or (2,15,7,10), and finds the probability of making past the first round and probability of making it past the second round.

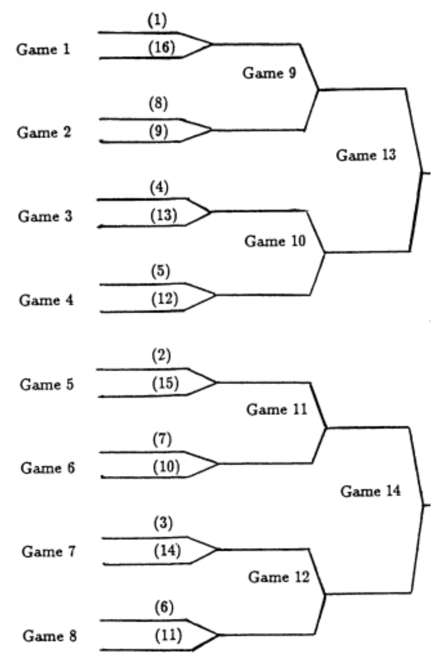

Figure 1 – March Madness Region Example

We can define $P_k(i,j)$ as the probability of a given team with seed i beating seed j in the k^{th} game of the region. Then we can define $P(i,j) = \frac{u(i)}{u(i)+u(j)}$ where, $i \neq j$ and,

$$u(x) = Performance\ Metric\ of\ x^{th}\ seed\ in\ region \text{ and } u(x) > 0$$

$$\therefore 0 < P(i,j) < 1$$

We can now see that if for example, we wanted the probability that the third seed made it to the second round, we can find it using $P_7(3,14) = P(3,14) = \frac{u(3)}{u(3)+u(14)}$. A Markov Chain matrix can now populated with $P(i,j)$ as the cells and the rows of this matrix will be seeds one through sixteen and the columns will be the same. From the matrix and the bracket we can see that if, for example, now we wanted the probability that the third seed made it to the third round, we can find it using,

$$P_{12}(3,j) = [P_7(3,14) * P_8(6,11) * P(3,6)] + [P_7(3,14) * P_8(11,6) * P(3,11)]$$

As an example, we will take the East region in the year 2011 and populate a table of the probabilities for each team to make it to round 2 and round 3 using the three methods define $u(x)$, the baseline seed method, ARIMA method, and exponential smoothing method.

Probabilites of Making Round 2 and 3 using Markov Chain

Team	Seed	Seed Method R2	Seed Method R3	ARIMA Method R2	ARIMA Method R3	ETS Method R2	ETS Method R3
Ohio St.	1	0.94	0.84	0.97	0.75	0.88	0.62
North Carolina	2	0.88	0.71	0.89	0.62	0.78	0.51
Syracuse	3	0.82	0.58	0.76	0.43	0.7	0.4
Kentucky	4	0.76	0.47	0.73	0.4	0.63	0.33
West Virginia	5	0.71	0.36	0.56	0.3	0.57	0.32
Xavier	6	0.65	0.26	0.74	0.41	0.55	0.29
Washington	7	0.59	0.17	0.59	0.24	0.56	0.24
George Mason	8	0.53	0.08	0.49	0.12	0.49	0.17
Villanova	9	0.47	0.06	0.51	0.13	0.51	0.18
Georgia	10	0.41	0.09	0.41	0.13	0.44	0.16
Marquette	11	0.35	0.1	0.26	0.08	0.45	0.21
Clemson	12	0.29	0.09	0.44	0.21	0.43	0.2
Princeton	13	0.24	0.08	0.27	0.08	0.37	0.15
Indiana St.	14	0.18	0.06	0.24	0.07	0.3	0.11
LIU Brooklyn	15	0.12	0.04	0.11	0.02	0.22	0.08
UTSA	16	0.06	0.02	0.03	0	0.12	0.03

Table 1 – Example of Different Methods Probabilies in East 2011 Region

We can see the differences between the three methods in this table, it is important to note that for any region the seed method will always have the same probabilities because the nature of $u(x)$ for that method is not dependent on the actual characteristics of the team. For this region and this year, we can see that the ARIMA method only predicts one upset, although it does predict a close first round game between 5 seed West Virginia and 12 seed Clemson. The ETS method predicts the same one upset as the ARIMA method but is much more favorable to the performance of the lower seeds in that region. This dynamic will change depending on the region and the year and this can be explored along with the performance of each method's ability to predict what happened.

Chapter 6

Analysis of Results

Sample

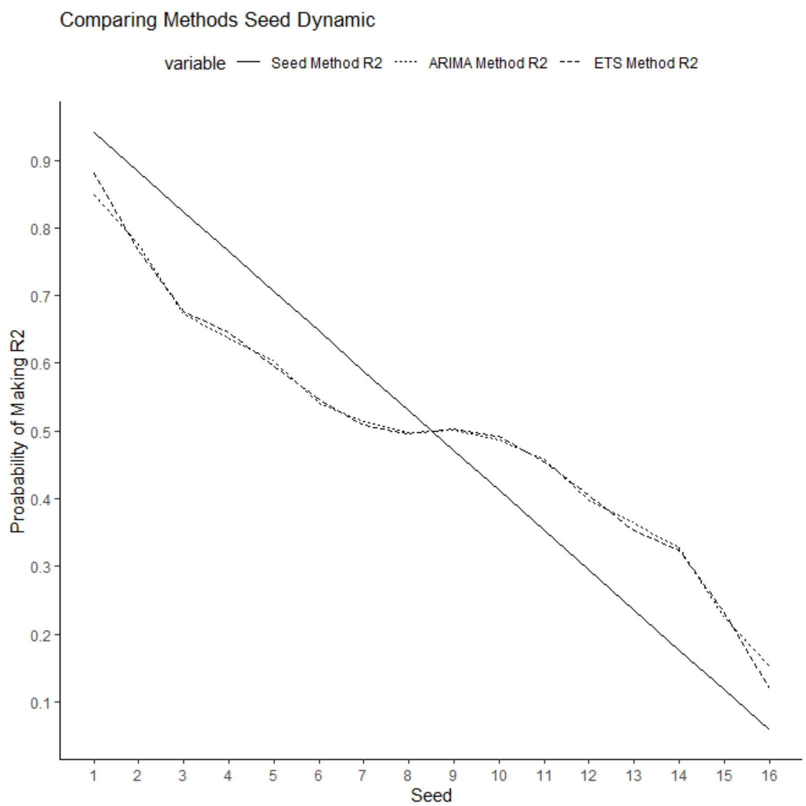

Figure 2 – Average Prediction Probability Curve

Chapter 7

Limitations and Conclusions

Sample

Appendix A

R Code

Dummy Code Dummy Code

BIBLIOGRAPHY

Sample Works Cited. Sample Works Cited. Sample Works Cited. Sample Works Cited. Sample Works Cited. Sample Works Cited. Sample Works Cited. Sample Works Cited. Sample Works Cited.

Sample Works Cited. Sample Works Cited. Sample Works Cited. Sample Works Cited.

Sample Works Cited. Sample Works Cited. Sample Works Cited. Sample Works Cited.

ACADEMIC VITA

Christopher Miller

Education

Pennsylvania State University, Schreyer Honors College University Park, PA
B.S. Mathematics, B.S. Applied Statistics, Minor Economics Expected May 2023

Awards

- Schreyer Honors College Scholarship
- Matthew Rosenshine Fund for Excellence in Statistics
- Phi Beta Kappa – National Honors Society
- Rensselaer Medal Award (Math and Science Excellence)
- NYS Scholarship for Academic Excellence
- Dean's List

Relevant Experience

MetLife Virtual - Bridgewater, NJ
Actuarial Intern (Disability Pricing) May 2022-Current
- Assisted with multiple projects that involved analysis with Excel VBA code, Python scripts, SQL, Alteryx, and R
- Improved disability claim predictive model in R by increasing the effectiveness of model code and helping decision making for factor analysis to create a model that improves pricing rates

Ginsberg's Foods Hudson, NY
Data Analyst Intern May 2021-March 2022
- Worked for a mid-size regional food distributor and trained in a data analyst role
- Lead longer-term projects by providing insight to the company such as operations utilization tools and zone pricing analysis and logic
- Investigated KPI vs. Distance Metrics using R Markdown while collaborating with management such as the CEO, COO, and VP of Operations

Extracurriculars

Esports Varsity Team (Division Head) Aug 2019-Current
- Lead a division where I held bi-weekly meetings with members, worked on creating tournament opportunities for members and moderated the community within the division

Actuarial Science Club Aug 2019-Dec 2022
- Former Executive Board member of the club, also a mentor for a student