ESZTERHÁZY KÁROLY CATHOLIC UNIVERSITY

INSTITUTE OF MATHEMATICS AND INFORMATICS

# Web Scraping and Monte Carlo Simulations for Analytical Forecasting

| **Author** | **Supervisor** |
|---|---|
| Lorand Heidrich | Adam Kovacs |
| Computer Science BSc | Teaching Assistant |

EGER, 2024

# Acknowledgments

I would like to express my gratitude to Dr. Eric Grono and Mr. Garrett Weinzierl for their mentorship and support throughout my journey into the domain of Monte Carlo simulation and analytical forecasting. Their expertise, patience, and guidance have not only paved the way for new avenues of exploration but have also instilled within me an appreciation for the intersection of computer science and predictive modeling.

Their influence has played an instrumental role in shaping my academic and professional trajectory. I am indebted to them for their impact on my endeavors.

Thank you both for your generosity in sharing your time and knowledge, and your patience in addressing my myriad of questions throughout my studies.

# Contents

# Chapter 1

# Introduction

In today's world, data has become of paramount importance, profoundly influencing our lives and shaping decision making processes. The acquisition, processing, and interpretation of data is fundamental across multiple domains. [1] Recognized as the cornerstone of contemporary insights, data serves as the basis of deriving valuable insights, and making informed projections, thereby guiding strategic planning and allowing for suitable preparation in the face of uncertainty. However, utilizing the full potential of acquired information effectively in a complex, multi-variable dynamic environment can be a challenging task [2].

This thesis approaches data collection and forecasting from a sports analytical perspective, aiming to derive statistical insights and formulate projections regarding future performance. It endeavors to utilize a combination of web scarping techniques [3] and Monte Carlo simulation [4] for analytical forecasting. Through the integration of these techniques, this research aims to explore a comprehensive methodology for data acquisition and predictive modeling.

## 1.1 Contextual Background:

The National Basketball Association (NBA) [5] is well known for its worldwide prominence and dedicated fan base. Its enduring popularity has resulted in a multitude of analytical data relating to historic games. This abundance of statistical data, along with a widespread general awareness of the sport and my personal enthusiasm for it, positions historic NBA games an ideal domain for exploring predictive modeling based on data obtained through web scraping.

## 1.2 Motivation:

The incentive for this research is derived from a keen interest in the technical intricacies of web scraping and probabilistic elegance of Monte Carlo simulations. The application of these techniques transcends the domain of sports analytics, with uses in finance [6], physics [4], and beyond [7].

## 1.3 Objectives:

The primary objective of this thesis is two-fold. Initially, to employ web scarping techniques to gather comprehensive historical NBA game data from the early 1990s. Subsequently, to utilize said data to simulate a general probabilistic outcome for selected historic NBA games.

Specifically, the research aims to:

- Develop a multi-approach web scraping pipeline to gather comprehensive historical data for a given NBA season and team.
- Manage and store the acquired data.
- Implement a multi-epoch Monte Carlo simulation to model potential game outcomes based on the attained data through modeling offensive possessions.
- Evaluate the predictive accuracy and reliability of the proposed methodology through empirical testing and validation against actual historic game results.

Through these objectives, this thesis undertakes to promote a deeper understanding of web scraping and predictive modeling within sports analytical forecasting.

# Chapter 2

# Methodology

## 2.1 Web Scraping Techniques

## 2.2 Monte Carlo Simulation

## 2.3 Utilized technologies

# Chapter 3

# Requirements

# Chapter 4

# Architecture

# Chapter 5

# Implementation

# Chapter 6

# Testing and Validation

## 6.1 Unit Testing

## 6.2 Integration Testing

## 6.3 System Testing

## 6.4 Performance Evaluation

# Chapter 7

# Results and Discussion

## 7.1 Analysis of Web Scraping Results

## 7.2 Evaluation of Monte Carlo Simulations

## 7.3 Comparison with Existing Methods

# Chapter 8

# Conclusion

## 8.1 Summary of Findings

## 8.2 Contributions to Knowledge

## 8.3 Limitations and Future Work

# Chapter 9

## 9.1

### 9.1.1

# Chapter 10

# Chapter title

## 10.1 Section title

### 10.1.1 Subsection title

Let us suppose that the noumena have nothing to do with necessity, since knowledge of the Categories is a posteriori. Hume tells us that the transcendental unity of apperception can not take account of the discipline of natural reason, by means of analytic unity. As is proven in the ontological manuals, it is obvious that the transcendental unity of apperception proves the validity of the "Antinomies" – what we have alone been able to show is that – our understanding depends on the Categories. [9, p. 102]

It remains a mystery why the Ideal stands in need of reason. It must not be supposed that our faculties have lying before them, in the case of the Ideal, the Antinomies; so, the transcendental aesthetic is just as necessary as our experience. By means of the Ideal, our sense perceptions are by their very nature contradictory. [9, 10]

**Theorem 10.1.** *Text.*

*Proof.* Text. □

**Definition 10.2.** Text.

*Remark* 10.3. Text.

# Chapter 11

# Appendices

## 11.1 Code Samples

## 11.2 GUI Mockups

## 11.3 Test Cases

# Bibliography

[1] Medium, *The Power of Data: Understanding Its Impact and Applications Across Various Domains*, Jonathan Mondaut, 2023, https://medium.com/@jonathanmondaut/the-power-of-data-understanding-its-impact-and-applications-across-various-doma [Retrieved: 2 March 2024]

[2] Northeastern University, College of Science, *Why it's so hard to make accurate predictions*, Jason Kornwitz, 2017, https://cos.northeastern.edu/news/hard-make-accurate-predictions/, [Retrieved: 27 February 2024]

[3] Moaiad Ahmad Khder, *Web Scraping or Web Crawling: State of Art, Techniques, Approaches and Application*, International Journal of Advance Soft Computing and Applications, Vol. 13, No. 3, 2021, Print ISSN: 2710-1274, Online ISSN: 2074-8523, Al-Zaytoonah University of Jordan

[4] Adekitan Aderibigbe, *A Term Paper on Monte Carlo Analysis/Simulation*, Department of Electrical and Electronic Engineering, Faculty of Technology, University of Ibadan, 2014.

[5] Wikipedia, *National Basketball Association*, 2024, https://en.wikipedia.org/wiki/National_Basketball_Association, [Retrieved: 27 February 2024]

[6] Don L. McLeish: *Monte Carlo Simulation and Finance*, Hoboken, New Jersey, USA, John Wiley & Sons, Inc., 2005.

[7] Paul Steffen: *Statistical Modeling of Event Probabilities Subject to Sports Bets: Theory and Applications to Soccer, Tennis, and Basketball*, Statistics [math.ST], Université de Bordeaux, 2022. English. NNT: 2022BORD0210. tel-03891393.

[8] , ,

[9] Donald Ervin Knuth: *Deformation modelling tracking animation and applications*, Berlin, Heidelberg, Springer, 2001.

[10] Christopher Manning, Prabhakar Raghavan, Hinrich Schütze: *Introduction to Information Retrieval*, New York, USA, Cambridge University Press, 2008.