Institute of Mathematics and Informatics

Web Scraping and Monte Carlo Simulations for Analytical Forecasting

Author
Lorand Heidrich
Computer Science BSc

Supervisor
Adam Kovacs
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

Chapter 3

Requirements

3.1 Requirements List

The system shall be constructed to uniquely fulfill both the requirements of a computer science thesis, and the domain of data acquisition and simulation based projections. It should therefore result in an intuitive end-user experience, leveraging the web scraping and Monte Carlo simulation methodologies explored in this thesis.

The client interface should allow users to interact with the business logic[1], thereby accessing the database through built-in functions. It should also allow for the utilization of web scraping and Monte Carlo methodologies. The Use Case diagram depicted below outlines the basic functionality described by the Functional - (see table 3.1), Non-Functional - (see table 3.2), and Platform Requirements (see table 3.3) outlined in this chapter.

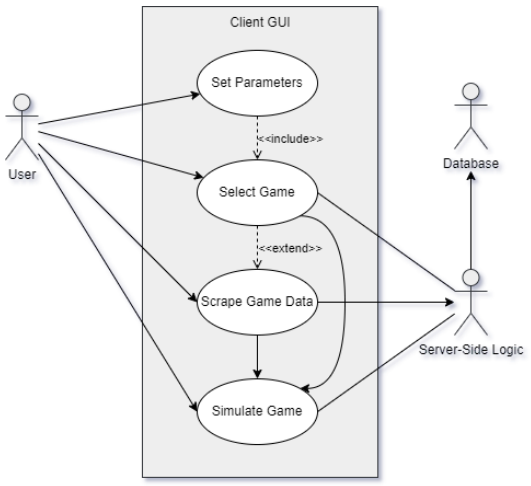

Figure 3.1: Use Case Diagram

[1]The term refers to the collection of algorithms responsible for allocating and processing data through communication with the database in order to serve the user interface, while maintaining its independence from both. For further information, please see [8].

3.1.1 Functional Requirements

ID	Name	Description
R1	Database	The system must allow users to select either the default database or utilize their own, based on a URI connection string.
R2	Game Parameters	It should provide users with a method to set the season, home- and away team.
R3	Epochs	The system must enable users to specify the number of epochs for the Monte Carlo simulation.
R4	Game Data	Historic game data should be displayed based on these settings for user review.
R5	Select Game	Users must be able to select an exact game to simulate from the displayed list of historic games.
R6	Missing Game	The system must recognize if the selected historic game is not in the database.
R7	Scrape Method	It should provide users with options for scraping the missing data trough different web scraping methods.
R8	Proxies	Scraping options should include the ability to use proxies.
R9	Proxy List	Users should have the ability to utilize their own proxy lists.
R10	Forced Scrape	The system must allow users the option to scrape game data even when it is deemed unnecessary by the algorithm.
R11	Validation	The system must ensure that data is not duplicated in the database.
R12	Simulation	The system must execute Monte Carlo simulations based on the selected game parameters.
R13	Graphs	It should visualize simulation results with graphs, including a probability density graph and a violin graph.
R14	Metrics	The system must return basic metrics such as the number of wins for each team and the mode of scores.
R15	Comparison	Users should be able to compare simulation results with original game data.

Table 3.1: List of Functional Requirements

3.1.2 Non-Functional Requirements

ID	Name	Description
NR1	Anonymity	The system must take steps to attempt anonymity throughout the web scraping process.
NR2	Validation	It should validate user input parameters, throwing errors when incorrectly set.
NR3	Errors	Users should be notified of errors during the application's operation.
NR4	Logging	It must utilize a logging system to allow for easier debugging.
NR5	Intuitive	The system must have an easy-to-use and intuitive interface.
NR6	Requests	The client side of the system must communicate with the server-side logic using HTTP to attain services as a responses.
NR7	SQL	The server-side logic should interact with the database using SQL queries.
NR8	Database	The system must be able to utilize separate MySQL database servers.
NR9	Testing	It should undergo thorough testing and validation to ensure accuracy, reliability, and robustness.
NR10	Regulation	The system must comply with relevant legal and regulatory requirements.

Table 3.2: List of Non-Functional Requirements

3.1.3 Platform Requirements

ID	Component	Requirement
PR1	Client	The application should be compatible with Windows 10 (or later) operating systems.
PR2	Client	The operating system is required to have .Net Framework 4.0.3 (or later).
PR3	Host	Server environment must be capable of running a Python application with a Flask framework.
PR4	Requirements	Back-end application requirements are available at: `https://github.com/lesheidrich/WebScraping_and_MCSim/blob/master/requirements.txt`.
PR5	Database	The database server must be compatible with either XAMPP or MySQL.

Table 3.3: List of Platform Requirements

3.1.4 Use Cases

- **Game Selection**: The user selects parameters such as the desired season, home- and away team, to initialize game selection, then chooses the desired match from the returned table.
- **Validation**: After accidentally setting a team to play against themselves, the user receives an error message alerting them of the mistake.
- **Scraping**: Following the game selection process, the system determines the game data is not in the database, then proceeds to utilize web scraping techniques to gather the data from online sources.
- **Simulation**: A user parameterizes the number of epochs for the Monte Carlo simulation and initializes game selection. The system utilizes the acquired historical data to run simulations, generating probabilistic outcomes for the historic NBA game.
- **Comparison**: Users compare the results of the Monte Carlo simulation with the original game data, assessing the accuracy of the model. The system further provides probability density- and violin graphs to further facilitate result analysis.
- **Error Handling**: When the user tries to initialize game selection, the database is down. The host service returns an error, notifying the user of the access issue. The user escalates the error, and upon its resolution normal system operations resume.

Chapter 4

Architecture

4.1 Design Concepts

At its core, the application relies heavily on a three principal layer [9, p. 19] concept, commonly found in systems utilizing a database and presentation layer. Fowler refers to these as presentation logic, domain logic, and data source. The presentation logic facilitates user interaction with the system, the data source handles data transactions and houses application information, while the domain logic's algorithms are responsible for data modification and layer interaction.

This is in line with the Gang-of-Four's [1] Model-View-Controller (MVC) [10, p. 529] design pattern. The View receives user-initiated interactions along with their parameters, and presents the application's data. It is capable of connecting directly to the model, while operating in conjunction with the Controller. The Controller interacts with both components as it processes their data and coordinates operations. The Model houses and manages the application data.

As discussed by Ahlan, A. R., Ahrnud, M. B., and Arshad, Y. [12], there have been several uses and variations of thin client applications since the 1970s. As a generalization, the *view* in its capacity as the client receives application data and logic based services from a host system. This application adheres to the this concept quite strictly, with the client acting as an intermediary between the user and the host, taking use parameters and displaying host response results. The host service, operating as the back-end, encompasses the previously discussed Model, Controller and all other components of the application.

[1] The Gang of Four [11] (GOF) are a group of four writers, all computer science professionals and entrepreneurs. Their literature and courses focus on professional development in the domain of computer science.

4.2 Components

Following the MVC design pattern's component structure, the application's presentation logic is allocated to the *view* package. The database and simple business logic allowing for record management is stored in the *model*. Acting as an intermediary between the two, the *controller* package orchestrates the flow of information along with its processing. The *webscraper* and *simulator* packages also tie into the *controller*, offering web scraping, and Monte Carlo simulation logic respectively, while further decoupling the application's components and allowing for better organization and maintainability through separation of concerns [2].

The application is organized in a manner, that all components embody system packages allowing improved readability and usability. The following subsections discuss each package's purpose and functionality.

4.2.1 *controller*

Purpose: The package is responsible for processing component interaction, and serves as the core logic of the system.

Functionality: Serving as the system's host, the *controller* is responsible for handling user prompts sent by the client in order to formulate an appropriate response. In order to fulfill this function it utilizes its connection to each component of the application, accessing their functionality as needed.

One of its responsibilities is accumulating data through the application of various web scraping methods, which are stored for future use. It's web scraper control functionality enables it to utilize the *webscraper* package to appropriate preset website data into memory, then reallocate it to the database using a combination of its own control processes along with the business logic of the *model* package. The web scraping sequence diagram illustrates the process (see 4.4).

When running a simulation for a parameterized game, the *controller* restructures relevant data for the selected historic games from the *model* into player, roster, and team objects usable by the *simulator*. The simulator returns the results to the host, which are forwarded back to the client. The Monte Carlo simulation sequence diagram shows further details (see 4.3).

Interaction: In its capacity as the main communications hub of the application, the *controller* interacts with every component in the system. The package's component

[2]Separation of concerns (SoC) is a software development design principle promoting segregation of source code elements by functionality in order to improve readability, organization and modification [13].

diagram provides a high level overview of basic component interaction (see4.1).

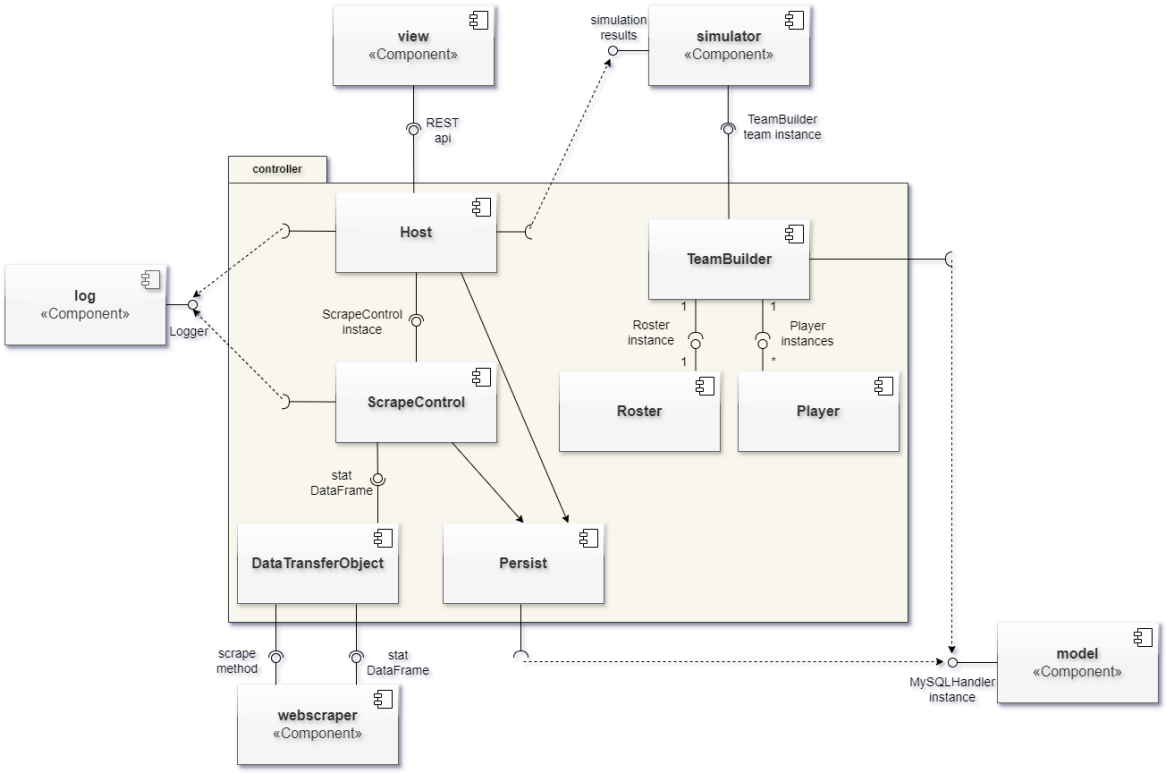

Figure 4.1: *controller* UML Component Diagram

4.2.2 *log*

Purpose: The component provides logging functionality for back-end operations.

Functionality: The *log* package contains logic allowing components to access system logs to document runtime errors and operations, useful for debugging potential issues. Log management functionality is also provided by the package, ensuring proper settings, size limitations and functionality.

Interaction: The *log* interacts with the *controller* and *webscraper* packages.

4.2.3 *model*

Purpose: The package ensures successful data management services within the application, through the storage and manipulation of data records and structures.

Functionality: The business logic enables system interaction with the database, encapsulating data access and administration services. Data security is enforced by minimizing vulnerabilities and prevents data corruption trough validation logic. The *model* services the system through the *controller* package's data manipulation and retrieval components.

The package also contains the NBA team enums [3] employed by the back-end logic. These ensure data validation and encapsulate all occurrences of team names in the system and its dataset, streamlining their application throughout processes.

Interaction: The primary business logic elements of the *model* communicate exclusively with the *controller*, enhancing system modularity (see 4.1). Due to their earlier described functionality, NBA team enums are also widely employed throughout the *simulation* package.

4.2.4 *simulation*

Purpose: The package's primary objective encompasses the repetitive simulation of a basketball game between selected teams at a specific point in time, with the end goal of returning the outcomes as graphical representations, comparable with the original historic game's outcome.

Functionality: The core functionality of the application revolves around implementing Monte Carlo simulations over a preset range of epochs to determine the probabilistic outcome of a historic NBA game.

The *simulator* package receives the relevant compiled data from the *controller*, initializing the creation of the simulation. Upon successful completion of the simulation, probability density and violin graphs are returned to the *controller* along with minimal game statistic like total win percentage and mode of scores reached per team. A detailed description of the process is illustrated in the Monte Carlo simulation's sequence diagram (see 4.3).

Interaction: The *simulator* relies on the *model*'s NBA team enum during operation, along with the *controller* for providing the necessary historic game data for simulation's successful operations. It further communicates with the *controller* in its capacity as the host. The *simulation* component diagram offers an overview of component interaction (see 4.2).

[3] In their capacity as a distinct object type, enumerations (enums) offer value binding across a group of encapsulated constants tied together through enumeration. Each value can act as a key identifier during instantiation, making enums a powerful structure for housing validated data [15].

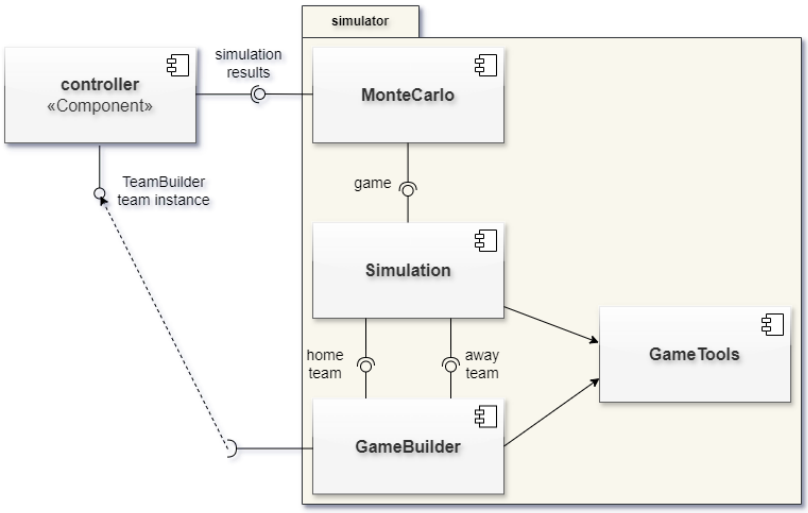

Figure 4.2: *simulation* UML Component Diagram

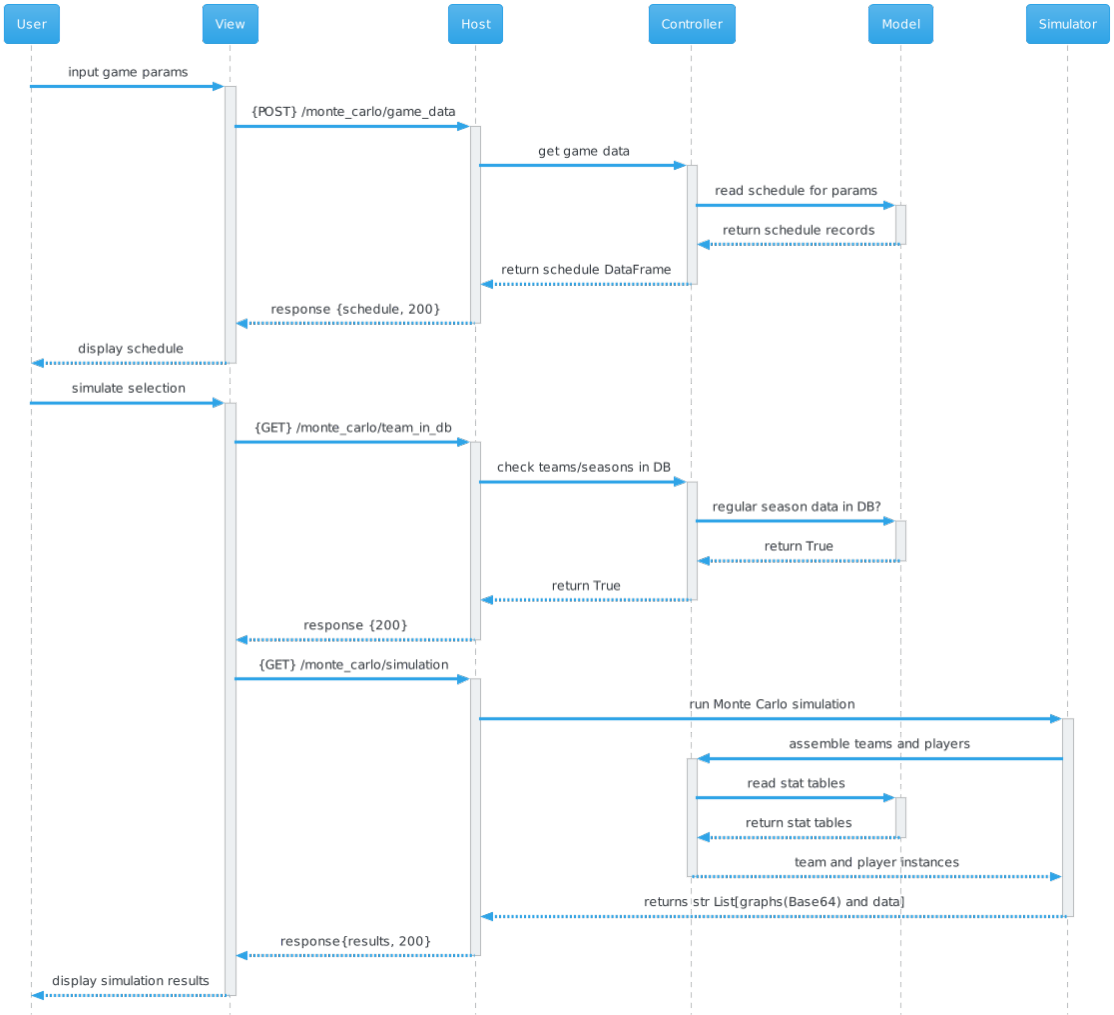

Figure 4.3: Monte Carlo Simulation UML Sequence Diagram

4.2.5 *test*

Purpose: The component is responsible for providing comprehensive back-end testing services.

Functionality: The test package constitutes a wide spectrum of tests focusing on the back-end of the system. It provides full-scale uni tests organized by module. Integration testing and linting is also included. The Testing and Validation chapter (see 6) contains a detailed breakdown of the package's functionality and implementation.

Interaction: Each unit test interacts with their respective components on the back-end. Integration tests interact with multiple packages following their functionality, in their attempt to check full system compatibility.

4.2.6 *view*

Purpose: To allow for user interaction with the system through communication with the host along with its application logic and data.

Functionality: The view package operates as a thin client, taking user input through a graphic user interface (GUI) [14]. User data is cached on the client side for the sake of user convenience, decreasing the input required to achieve functionality. Host responses, along with operational errors are displayed for user viewing. Errors are not logged on the client side of the application.

Interaction: The component interacts with the user and the host module of the *controller* package.

4.2.7 *webscraper*

Purpose: The package completes data gathering services from the amalgamation of preset websites and parameters representing selected NBA games, in order to supply historic statistical game data for the application.

Functionality: The acquisition process extracts data from a combination of preset URLs, set to match parameterized game data originating from the user. It includes functionality to interact with each URL in a manner defined by the user's chosen scraping method. Collected information is preprocessed and transformed before being committed to memory in a preset reusable format, easily accessed by the *controller* as it looks to the *webscraper* to acquire new information. Web scraping requests are generally not repeated, as the system is built to house already accessed data, thereby

minimizing dependency on online sources to bare necessity. The component's sequence diagram (see 4.4) illustrates the process during runtime.

Interaction: The *webscraper* package primarily communicates with the *controller* in order to receive data acquisition requests and parameters. Its preprocessed results are also returned to the *controller*. The package's component diagram (see 4.5) illustrates the *webscraper*'s interactions. Logging is utilized throughout the process.

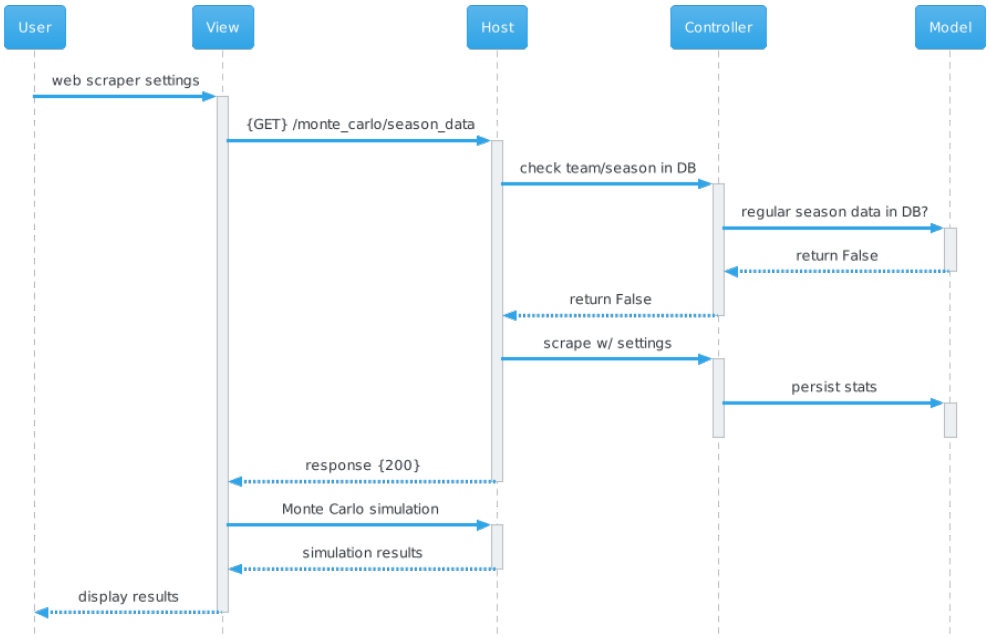

Figure 4.4: Web Scraping UML Sequence Diagram

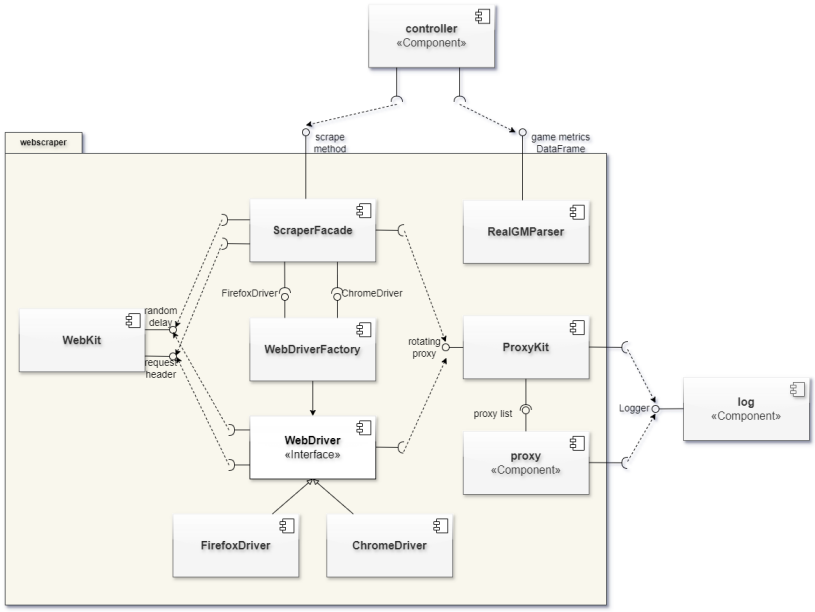

Figure 4.5: *webscraper* UML Component Diagram

4.3 Technologies and Frameworks

4.3.1 .NET

The .NET framework was built by Microsoft for developing Windows-based applications. Today, with the integration of cross-platform and open source frameworks, the .NET Core supports multiple operating systems and is also open source. It supports multiple programming languages, including C#. [36]

The view component of the application is based on the .Net framework and written in C#, due to its ability to model a range of services. It's WinForms (see 4.3.15) framework and RestSharp package allow for fast and secure development, as well as a seamless user exprience.

4.3.2 Beautiful Soup 4

Beautiful Soup [17] is a popular python package, designed to extracts data from HTML [4] and XML [5] documents sourced from the web. By leveraging attributes unique to the selected markup language, it facilitates tag and text content based parsing in order to precisely identify and extract the desired data. Developed for web scarping purposes, it is often used alongside packages responsible for making content requests to websites.

This was also the case in this project. Beautiful Soup was a great asset during the allocation and extraction of acquired web contents, which could then be passed on for further formatting and storage.

4.3.3 Fake User Agent

Throughout the extraction of data from websites it is crucial to consider the exchange of information between the client making the request, and the website's host server responding to the request. During this process, the host receives client information in the header, allowing it to discern information about the requesting client, like its operating system and browser type (see subsection 4.3.5 for information on HTTP requests and their headers).

Operations requiring repeated content requests to a single host, such as web scraping, can therefore be identified as outside the scope of regular usage requirements provided by the website. This in turn may lead to limitations of service, disrupting web scraping-based system operations.

[4]HyperText Markup Language (HTML) is a markup language used to house online content and set its layout, thereby creating the basic structure of web pages [16].

[5]Extensible Markup Language (XML) is a markup language for data transfer and storage. Rather than offer preset tags, users can organize content subjectively through the application of a key value pair data structure [18].

Fake user agent applications are designed to solve this problem. Python's *fake-useragent* library allows for the creation of random user agent headers, which can be assigned to requests [20]. Users have the option to select from a wide range of preset options, which can also be filtered to only mimic specific browser technologies and operating systems. While this method only changes a portion of the request data, when used alongside other methodologies promoting anonymity, it can prove to be a powerful tool.

The package was employed in the project for this ability to generate random user agents at runtime, thereby contributing to the application's anonymity while web scraping.

4.3.4 Flask

Flask is a Python based web framework created to facilitate the development of web applications. The framework is known for its versatility, and aims to provide a minimalistic and flexible approach in comparison with other web application frameworks [26].

Developers can choose their server platforms and environments, as Flask's WSGI (Web Server Gateway Interface) specifications enable it to integrate with a multitude of web servers, such as Apache and Nginx [25].

Flask also contains a built-in development server, which provides application testing and debugging features. It also supports other extensions and libraries, granting integration capabilities to developers to handle session management or database connectivity. The framework focuses on the provision of necessary tools instead of enforcing constraints, thereby making Flask a popular web framework [25].

The framework was employed in the application due in part to its ease-of-use and flexibility. The host required a lightweight solution compatible with Python. A robust strict framework such as Django would have hindered the timely development of the project, and Flask's accessibility made it a great candidate.

4.3.5 HTTP

Hypertext Transfer Protocol (HTTP) is a method of communication employed to facilitate data exchange between servers hosting web-based resources, and clients such as web browsers or web-applications [19].

HTTP is unidirectional, operating on a request-response basis. When the client, such as a web browser requests services in the form of online content from a server, it does so through an HTTP request. The server processes the request, and sends an appropriate HTTP response. Data can be transmitted in multiple formats. Common

examples include JSON [6] , HTML, and XML [23].

The client's request contains information such as the request line, message body, and headers [21]. The request line contains the destination resource, the message body stores parameters used by request methods, and the headers contain information about the client's choice of operating system and web browser. The response is comprised of a status code, headers and message body, which houses the requested data [23].

HTTP offers methods which can apply parameterized actions to the selected resource. Some examples include:

- The GET method presents the specified resource.
- POST submits data parameters to the server which are required for further functionality.
- PUT is used to update a selected resource on the server.
- The DELETE method removes the parameterized resource from the server.

Status codes are utilized to indicate the outcome of processes initiated by the client's HTTP request. They are categorized into groups, covering categories such as informational responses of success as well as errors. Some common examples of success codes include:

- 200: The requested transaction was completed successfully and the requested content is returned.
- 204: The server processed the request, and is not returning any content.
- 400: A client error obstructed the server's attempt to complete the request.
- 404: The server does not contain the requested resource.
- 500: Indicates the occurrence of an internal server error, obstructing the fulfillment of the client request.

Due to security concerns, a more secure version of HTTP Secure (HTTPS) [24] was created. It employs encryption mechanisms to ensure transaction security, thereby thwarting potential eavesdropping and tampering attempts.

[6] JavaScript Object Notation (JSON) is a popular lightweigth data structure which utilizes key-value pair functionality. Due to it's simplicity and compatibility with many programming languages it is a common choice for online information exchange [22].

4.3.6 Matplotlib

Matplotlib is an extensive community-maintained Python library utilized in the creation of static, animated, and interactive visualizations [27]. Matplotlib offers a wide range of functionalities, making it a versatile tool for data visualization tasks. It's Pyplot module allows for the creation of customizable and embeddable graphs and diagrams.

Ultimately the plotting capabilities of the package's Pyplot module led to its utilization in the system.

4.3.7 MySQL

MySQL is a free open source relational database management system (RDBMS). It is widely used due to it's speed, reliability and scalability. Structured Query Language (SQL) is utilized to communicate with the database for data management and user access modifications. [28]

The decision to use MySQL as the application's database management system was reached due to several factors. The application required a relational model for easier storage of acquired statistical data, which in turn allowed easier processing upon extraction from the database. Its reliability and complementary open source nature, coupled with its widespread adoption and accessibility solidified the decision to utilize it in the application.

4.3.8 Pandas

The pandas library is an open source project developed for Python by Wes McKinney and Chang She during their time at AQR Capital Management. The developers sought to attain the tabular functionality of DataFrames [7] in the R programming language for their flexibility and functionality in working with financial data. [29]

Pandas remains an open source library to this day, and has become very popular for its versatility and functionality in data analysis endeavors. It has been employed in this project due to this high level of tabular functionality and data accessibility, along with its efficiency in parsing tabular data into DataFrames.

4.3.9 Python

Python is a high level open source object-oriented programming language, with features such as dynamic typing, dynamic binding, and built in data structures. It is an

[7] DataFrames are two-dimensional tabular data structures in pandas, housing data accessible by rows and columns. Visually, DataFrames resemble spreadsheets, while structurally they key-value pair data structures. [29]

interpreted language with a modular structure, and an easily readable syntax. [30]

The programming language was utilized for the back-end due to its extensive complementary open source library. There availability of multiple web scraping packages, along with powerful data analytics libraries such as pandas or matplotlib allow for convenient and fast paced development.

4.3.10 Requests

The Python requests library is a simple package allowing for communication through HTTP requests. [31] It is utilized in the application for making requests to specified websites in order to retrieve their HTML content in the response. Requests allows for parameterized customization of proxies, timeout settings, and request methods, many of which are utilized in the web scraping process.

4.3.11 REST API

Representational State Transfer (REST) refers to a set of principles acting as guidelines in the development of APIs for web services. These stipulate the use of HTTP as a communication protocol, however data encoding is left up to the developer, with options including JSON, HTML, and XML. Requests should be independent, not relying on each-other's success. Caching of data is acceptable, along with the sending of executable code to the client where needed. REST further requires a standard method of sending data and a layered organizational approach. [33]

In the application, REST is utilized to facilitate communication between the client and host. HTTP requests are sent to the host in order to receive responses in the form of JSON.

4.3.12 RestSharp

RestSharp is a tool utilized in .NET development, offering synchronous and asynchronous communication to remote resources using HTTP. It allows for easier managing of diverse request and response types while interacting with APIs by handling serialization and deserialization of message bodies to JSON and XML. [32]

The package is utilized on the client side of the application, due to these capabilities. Its GET and POST methods handle interactions with the host's API. Response JSONs are deserialized to access response content.

4.3.13 Selenium

Selenium is an open source test automation tool created for web applications. It enables developers to single out and interact with user interface (UI) elements thereby

simulating user interaction. Selenium's WebDriver tool allows for the utilization of web browsers to interact with online sources, enabling programmatic clicking of buttons, mouse movements, and traversing web pages. Selenium supports multiple programming languages, including Python. [34]

While driver instantiation already allows for the use of well known browsers, such as Mozilla's Firefox and Google's Chrome, the open source community has further contributed packages such as Undetected ChromeDriver. This Python library aims to provide the same driver functionality, while attempting to be less detectable by the host server. [35]

This project utilizes Python's Selenium package for web scraping purposes. Instantiating the driver allows for content retrieval in a less detectable, albeit slower manner.

4.3.14 TestStack.White

White is a test automation framework used to test Windows desktop applications. It is based on the .NET framework and does not require the use of scripting languages. Test code can be written in any .NET supported language. [38]

The framework is implemented in the project for UI automation testing purposes. The client's WinForms-based desktop application is tested for returning appropriate responses and basic functionality for quality assurance purposes.

4.3.15 WinForms

Windows Forms (WinForms) is a .NET graphical user interface (GUI) framework for building desktop applications. It provides developers with controls for dragging and dropping elements to rapidly create interactive user interfaces. [39]

The application's client interface is built with WinForms in the C# programming language. WinForms allowed for timely and precise development process, creating a visually appealing Windows desktop application.

4.3.16 XAMPP

Cross platform Apache MariaDB PHP Pearl (XAMPP) is a complementary open source web server solution stack meant for development environment utilization. Originally developed by Apache Friends, the application is widely used for testing and development purposes. It's available for multiple operating systems and makes the conversion process to a live server seamless, as it utilizes the same tech stack utilized by most production environments. [37]

MariaDB, the open source database server used by XAMPP, is a fork of the the original MySQL. It was created with the intention that the project remain free and open source. [40] As a fork of MySQL, it shares a high level of compatibility with it, and has been included in the project for this purpose. While XAMPP employs MariaDB, the project's business logic utilizes MySQL related code and Python libraries. This compatibility allows the application to access both MySQL servers and XAMPP's MariaDB-based service.

XAMPP is utilized as a development environment in the project, housing the application's built in database. It was chosen for its robust functionality and online user interface, which made testing during the creation and implementation of the model a more user-friendly experience.

Chapter 5

Implementation

In keeping with the structural approach of the Architecture chapter, the project's implementation is presented by package. Each section will delve deeper into the modular and functional breakdown of the project, along with their application. As other components require its presence, the *model* package is introduced first, along with an introduction of the business logic.

5.1 *model*

discuss database architecture too

5.2 *controller*

The package is responsible for the implementation of three high level functionalities: the API host service, web scraping, and constructing *TeamBuilder* NBA team instances for the simulator. Each of these controls manage a designated section of the back-end logic's operations.

As discussed in the Architecture chapter (see 4), the *Host* amalgamates final operations of both web scraping -, and Monte Carlo simulation services. Web scraping operations are coordinated by the *controller*'s *ScrapeControl* class, while NBA team instantiation is handled by the *TeamBuilder*. Team instances are provided for the *simulator* package, which in turn services the *Host*. The following sub-sections presents each functionality in detail.

5.2.1 Host Service

The *host_service.py* module houses the *Host* class containing the Flask server handling requests pertaining to the functionality provided by the above described *webscraper*

and *simulator* packages. Server instances handle these requests through the following endpoints and their respective class methods:

- **/monte_carlo/game_data**: Its get_game_data() method reads the game schedule from the database based on the provided arguments, returning a JSON of the records to the client.
- **/monte_carlo/team_in_db**: The get_teams_in_db() method checks if the selected and previous season's data is present in the database for both teams, notifying the client of its findings.
- **/monte_carlo/season_data**: Its get_season_data() method initiates a scrape of missing game data. In case of failure, the method re-initiates the process for a second time, logging the attempt. The client is notified of the operation's success or failure via a REST response. In case of failure, the method removes records from the players table, thereby ensuring future checks will see it is still missing from the database.
- **/monte_carlo/simulation**: The get_monte_carlo_sim() initiates the parameterized Monte Carlo simulation, returning its results to the client. The JSON contains the probability density plot and violin plots as base64-encoded strings of the plot image files created by matplotlib. Basic game statistics are also included, such as modes of the score arrays, and win percentage.

Flask instances are initialized with a timeout of 500 seconds, just over 8 min. allowing adequate time for slightly longer simulations (see line 5 in Code Extract 5.1).

Utilizing matplotlib to create graphs during server runtime was challenging at first. When the server instance was running alongside the WinForms GUI, threading conflicts would randomly occur. This was because matplotlib's default back-end GUIs are not guaranteed to be thread-safe, therefore WinForms would occasionally attempt to utilize a resource already allocated to the plotting function. The solution was to set matplotlib to the Anti-Grain Geometry (Agg) plotting library (see line 3 in Code Extract 5.1), which does not require the use of said resources. [41]

Code Extract 5.1: Flask Instance Initialization

```
1     def __init__(self):
2         # non-interactive rendering env
3         matplotlib.use('Agg')
4         self.app = Flask(__name__)
5         self.app.config['TIMEOUT'] = 500
6         self.log = Logger(log_file="application_log.log",
7                           name="FLASK HOST",
8                           log_level="INFO")
```

5.2.2 Web Scraping

The *controller* package's web scraping functionality is encapsulated in two modules: *control_service* and *dto_service*. Each module's classes contribute specialized logic to the data gathering process. Together, the modules orchestrate the web scraping process, handle the transformation of data in memory, and manage persistence operations through interactions with the business logic. Figure 5.1 illustrates the package's class diagram pertaining to web scraping and data allocation logic.

The *control_service* module houses the *ScrapeControl* class. During initialization it:

- Instantiates a *DataTransferObject* (see section 5.2.2) utilizing the specified proxy settings.
- Creates a dictionary of key-value pairs comprised of the parameterized URLs for each web page it intends to scrape data from.
- Assesses if the team is in the playoffs, storing a boolean from the result of the function call to *Persist.team_in_playoffs*.
- Initializes a logger instance, and sets the necessary attributes for further operations.

The class is comprised of basic logic for scraping each table type, stored in the database: player, individual game, and team statistics. These are utilized by two main operating functions: *run_single* and *run_all*. This is because the application is designed to allow for simple use through a client interface, but also complex back-end use for a more detailed control of the process as required by analysts. The *run_all* method completes a full scrape for the chosen team and season, while the *run_single* method takes parameters allowing for further specification of the table type to acquire and populate (player, individual game, or team statistics).

The dictionary data structure housing the URLs allows for a controlled and detailed iteration, utilizing the assessment of the team's participation in playoffs for the given season set up during the instance's initialization. While iterating over the URLs, the appropriate scrape methods are utilized to attain raw HTML data as text, parse it to a Pandas DataFrame, and finally persist the data to the database. This functionality is further discussed in section 5.7 titled *webscraper*.

The *dto_service* module contains the *DataTransferObject* and *Persist* classes. *DataTransferObject* instances facilitate the retrieval of data through *ScraperFacade* instances (see section 5.7). Users have the option to choose from a variety of scrape

methods, with or without the utilization of proxies. The class further provides logic for handling data sourced from each pre-determined URL, flipping through pages where they are available, and storing them in Pandas DataFrames.

Beyond utilizing services encapsulated by the *webscraper* package, the class aims to alter and store data in memory for easier processing and better accessibility. Utilizing Pandas DataFrames as a data transfer object [1] allows the class to compartmentalize data in structures that are easy to handle, debug, view and transfer to database tables.

With a collection of static utility methods, the *Persist* class is responsible for handling data management operations within the application's database. Trough interaction with the *MySQLHandler*'s business logic, it facilitates the insertion, deletion, and retrieval of data.

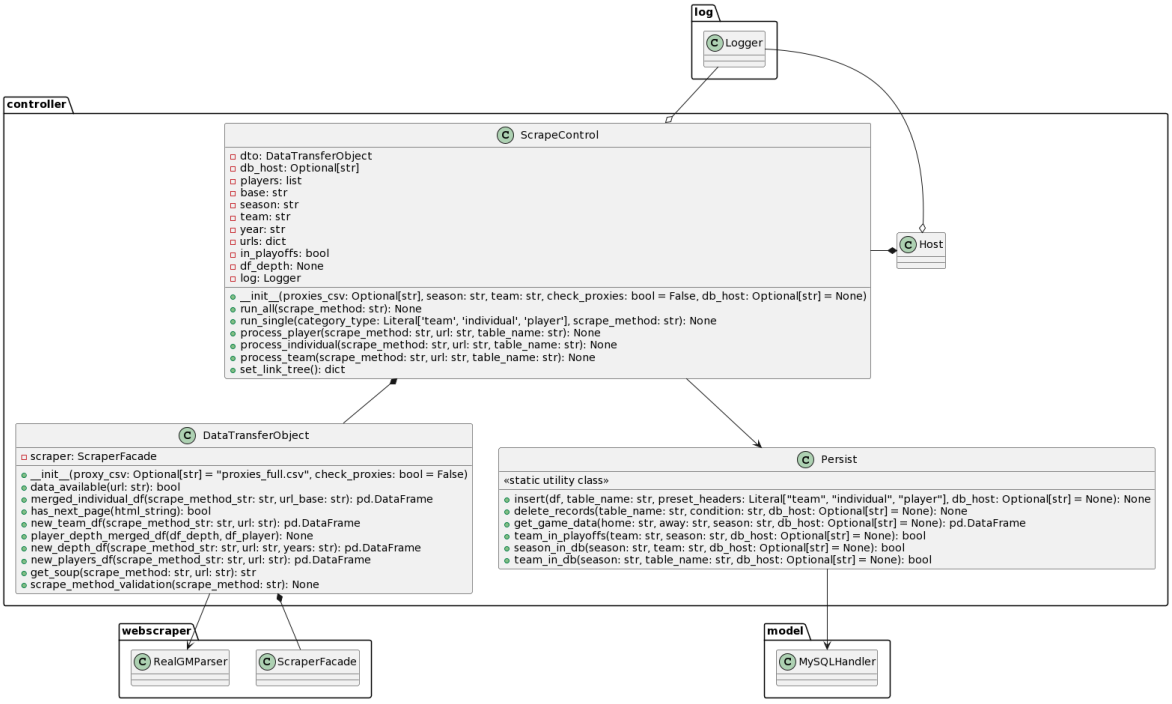

Figure 5.1: The *controller*'s Web Scraping Services UML Class Diagram

5.2.3 Team Instantiation

The *dto_sim* module contains three classes responsible for constructing NBA team instances: *TeamBuilder*, *Roster*, and *Player*. Completed team instances are utilized by the *simulator* package's *GameBuilder* class during the construction of individual game simulations.

[1] Data transfer objects (DTO) facilitate data transfer between processes, where the only behavior allowed to the data structure is the storage of data [42].

- see the graph for details - what attribs it has and why - how it packs players into roster - what player and roster is and how they work

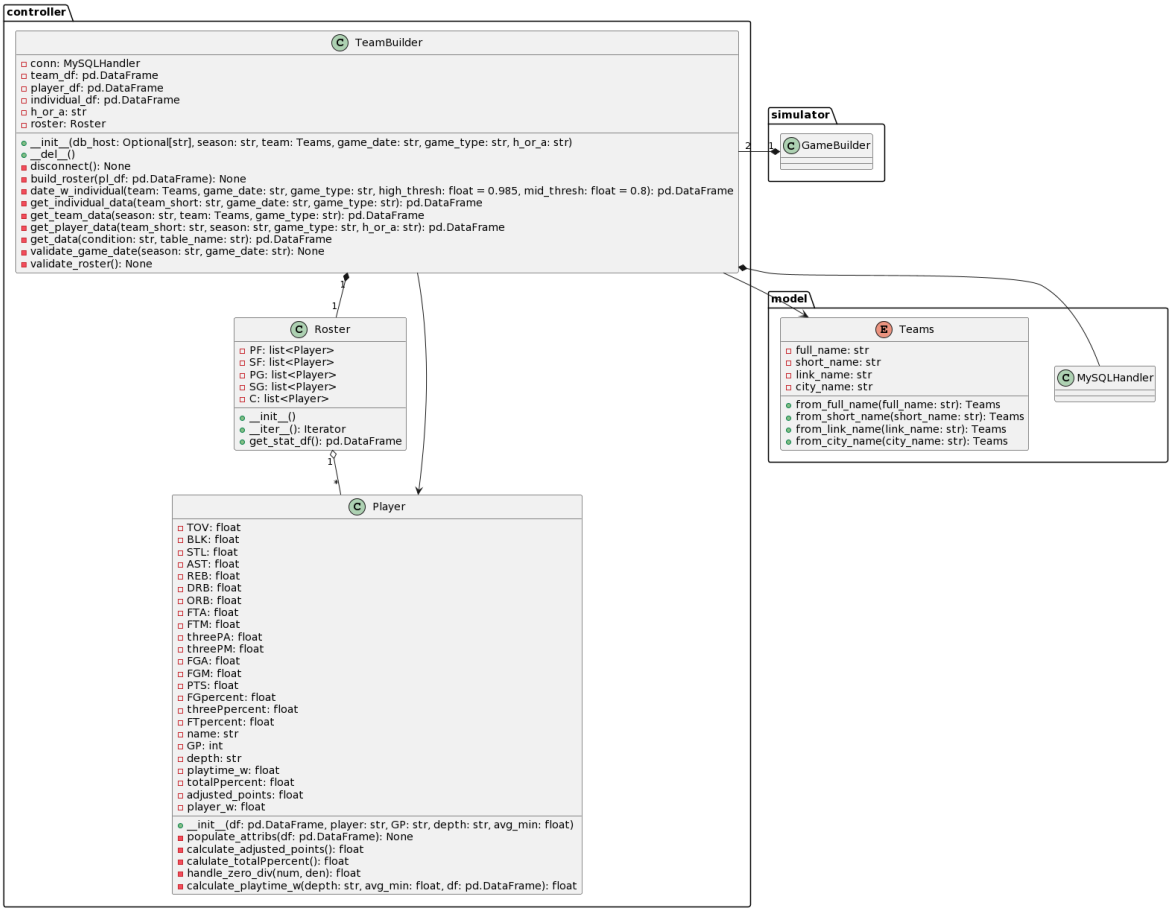

Figure 5.2: The *controller*'s TeamBuilder UML Class Diagram

These are accessed through a set of REST API-based methods organized into a class.

implementation approach actual code design patterns and standards challenges along the way and problems i ran into why'd you do it like that scalability

5.3 *log*

5.4 *model*

5.5 *simulation*

5.6 *view*

5.7 *webscraper*

Chapter 6

Testing and Validation

6.1 Unit Testing

6.2 Integration Testing

6.3 System Testing

with white

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

common to use scrapy. selenium is not preferred in general due to lack of speed. undetected chrome needs to be looked at further

Chapter 8

Conclusion

8.1 Summary of Findings

8.2 Contributions to Knowledge

8.3 Limitations and Future Work

Theorem 8.1. *Text.*

Proof. Text. □

Definition 8.2. "Antinomies"

Remark 8.3. Text.

Chapter 9

Appendices

9.1 Code Samples

9.2 GUI Mockups

9.3 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] Grady Booch, Robert A. Maksimchuk, Michael W. Engle, Bobbi J. Young, Jim Conallen, Kelli A. Houston, *Object-Oriented Analysis and Design with Applications*, Massachusetts, USA, Addison-Wesley, 2007.

[9] Martin Fowler, David Rice, Matthew Foemmel, Edward Hieatt, Robert Mee, Randy Stafford, *Patterns of Enterprise Application Architecture*, USA, Addison-Wesley Professional, 2002.

[10] Eric Freeman, Elisabeth Freeman, Bert Bates, Kathy Sierra, *Head First Design Patterns*, O'Reilly, 2004.

[11] Spring Framework Guru, *National Basketball Association*, 2024, `https://springframework.guru/gang-of-four-design-patterns/`, [Retrieved 5 March 2024]

[12] Abdul Rahman bin Ahlan, Murni bt Mahmud, Yusri bin Arshad, *Conceptual Architecture Design and Configuration of Thin Client System For Schools in Malaysia: A Pilot Project*, Department of Information System, Kulliyyah of Information and Communication Technology, Kuala Lumpur, Malaysia, 2010.

[13] Chris Reade, *Elements of Functional Programming*, Boston, USA, Addison-Wesley Longman, 1989.

[14] Wikipedia, *Graphical user interface*, 2024, `https://en.wikipedia.org/wiki/Graphical_user_interface`, [Retrieved 5 March 2024]

[15] Microsoft Learn, *Enumeration types (C# reference)*, Bill Wagner, 2023, `https://learn.microsoft.com/en-us/dotnet/csharp/language-reference/builtin-types/enum` [Retrieved 6 March 2024]

[16] MDN Web Docs, *HTML: HyperText Markup Language*, 2024, `https://developer.mozilla.org/en-US/docs/Web/HTML`, [Retrieved 6 March 2024]

[17] Beautiful Soup 4.12.0 documentation, *Beautiful Soup Documentation*, 2004-2023 Leonard Richardson, `https://www.crummy.com/software/BeautifulSoup/bs4/doc/#`, [Retrieved 6 March 2024]

[18] MDN Web Docs, *XML: Extensible Markup Language*, 2024, `https://developer.mozilla.org/en-US/docs/Web/XML`, [Retrieved 6 March 2024]

[19] MDN Web Docs, *HTTP*, 2024 `https://developer.mozilla.org/en-US/docs/Web/HTTP`, [Retrieved 6 March 2024]

[20] PyPI Python Package Index, *fake-useragent*, 2023, `https://pypi.org/project/fake-useragent/#description`, [Retrieved 6 March 2024]

[21] MDN Web Docs, *HTTP headers*, 2024, `https://developer.mozilla.org/en-US/docs/Web/HTTP/Headers`, [Retrieved 6 March 2024]

[22] MDN Web Docs, *JSON*, 2024, https://developer.mozilla.org/en-US/docs/Glossary/JSON, [Retrieved 7 March 2024]

[23] Wikipedia, *HTTP*, 2024, https://en.wikipedia.org/wiki/HTTP, [Retrieved 7 March 2024]

[24] Wikipedia, *HTTPS*, 2024, https://en.wikipedia.org/wiki/HTTPS, [Retrieved 7 March 2024]

[25] Pallets Projects, *Flask*, https://flask.palletsprojects.com/en/3.0.x/, [Retrieved 7 March 2024]

[26] Read the Docs, *Flask*, 2024, https://readthedocs.org/projects/flask/, [Retrieved 7 March 2024]

[27] Matplotlib, *Matplotlib: Visualization with Python*, 2023, https://matplotlib.org/, [Retrieved 7 March 2024]

[28] Wikipedia, *MySQL*, 2024, https://en.wikipedia.org/wiki/MySQL, [Retrieved 7 March 2024]

[29] Wikipedia, *pandas (software)*, 2024, https://en.wikipedia.org/wiki/Pandas_(software), [Retrieved 8 March 2024]

[30] Python, *What is Python? Executive Summary*, https://www.python.org/doc/essays/blurb/, [Retrieved 8 March 2024]

[31] Read the Docs, *Requests: HTTP for Humans*, https://requests.readthedocs.io/en/latest/, [Retrieved 8 March 2024]

[32] RestSharp, *Recommended usage*, Peter Breen, 2023, https://restsharp.dev/intro.html, [Retrieved 8 March 2024]

[33] Red Hat, *What is a REST API?*, 2020, https://www.redhat.com/en/topics/api/what-is-a-rest-api, [Retrieved 8 March 2024]

[34] Harvard Scholar, *Selenium Documentation Release 1.0*, 2012, https://scholar.harvard.edu/files/tcheng2/files/selenium_documentation_0.pdf, [Retrieved 8 March 2024]

[35] Github, *undetected-chromedriver*, 2024, https://github.com/ultrafunkamsterdam/undetected-chromedriver, [Retrieved 8 March 2024]

[36] Microsoft Learn, *Introduction to NET*, 2024, https://learn.microsoft.com/en-us/dotnet/core/introduction, [Retrieved 8 March 2024]

[37] Wikipedia, *XAMPP*, 2024, https://en.wikipedia.org/wiki/XAMPP, [Retrieved 8 March 2024]

[38] Read the Docs, *TestStack.White*, https://teststackwhite.readthedocs.io/en/latest/, [Retrieved 8 March 2024]

[39] Microsoft Learn, *Desktop Guide (Windows Forms .NET)*, 2023, https://learn.microsoft.com/en-us/dotnet/desktop/winforms/overview/?view=netdesktop-8.0, [Retrieved 8 March 2024]

[40] MariaDB Foundation, *About MariaDB Server*, https://mariadb.org/about/, [Retrieved 9 March 2024]

[41] matplotlib, *The builtin backends*, https://matplotlib.org/stable/users/explain/figure/backends.html, [Retrieved 10 March 2024]

[42] Wikipedia, *Data transfer object*, 2024, https://en.wikipedia.org/wiki/Data_transfer_object, [Retrieved 14 March 2024]

[43] , , , [Retrieved 14 March 2024]

[44] , , , [Retrieved 14 March 2024]

[45] , , , [Retrieved 14 March 2024]

[46] , , , [Retrieved 14 March 2024]

[47] , , , [Retrieved 14 March 2024]

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

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

tools: - draw.io - plantuml.com - chat.openai.com - stackoverflow.com - google scholar

ADD: - installation