Probability Analysis with Web Scraping and Linear Regression

Author: Javier A. Garcia Matos
Advisor: Dr. Nelliud Torres
Electrical and Computer Engineering and Computer Science Department

Abstract

Since the dawn of sports, probability and trends have played an important role in predicting the outcome of an event. It gives the public a general idea of how the match will occur, providing a powerful tool in analytics. This project intends to extract raw tennis match data from verified sources and apply mathematical equations to predict the probability of the outcome of a particular match. Firstly, with the help of the programming language Python and a popular technique known as web scraping, the data can be extracted from a verified source, such as the Association of Tennis Professionals (ATP), and validated. After the data is extracted, in this project, three different algorithms will be applied, with the goal of predicting the outcome of a particular tennis match. These algorithms are known as linear regression (decision tree, ridge, and lasso) and are made with the programming language Python.

Introduction

Since the sixteenth century, probability has played a large role in the world as we know it. From complex mathematical equations to a simple roll of the dice, the chance of an event occurring is present in everyday life. The world of tennis is not exempt from this. In every match, the odds are stacked against or for a player, depending on a variety of factors. In this project, we will gather the raw data from verified sources with web scraping and apply linear regression to understand the probabilities of a tennis match.

Background

To investigate the probabilities of a tennis match, some basic rules of the sport must be explained. First, the sport is played by either two or four players, with the same number on each team, that is, one player against another or two players against two. Players use rackets to hit the ball back and forth across the court. The points mainly come from three situations: when the ball gets stuck in the net, goes out of bounds, or bounces twice on the court; or when the opposing player does not hit the ball accordingly. The game is organized into sets, which are made up of games. Matches are typically played as a best-of-three or best-of-five sets, with each set consisting of games and potentially tiebreakers (different format games to decide the result of a set or match). Each set has a max total of ten games with the possibility of extending the games to 12 with a famous tiebreaker. In this setup, each game has a scoring pattern of 15, 30, 40, and 60 points. When a game gets to a tie at 40-40, a player needs to win by two points in a row.

Problem

With modern-day tennis and a variety of factors, the prediction of the outcome of a particular match has grown to be more difficult as time progresses. With newer techniques, players are surprising the outcome of tennis matches around the world. Consequently, a more effective solution to properly predict a match is needed. With this goal in mind, this project will concentrate on using those factors to produce an accurate prediction.

Methodology

In this section, the first step is to implement web scraping. Like a web API, web scraping provides a tool for interaction with a website or application, yet the main difference is in the approach used. An API is given by the developers of the application, while web scraping is stand-alone, that is, it extracts the data it sees in the front-end of the website. Consequently, with web scraping, tennis match historical data can be extracted based on tournaments or years, among other factors. To implement this, firstly, the verified source is found. Then, the following libraries are imported: Beautiful Soup (made for parsing the HTML), Selenium (webdriver to automate the front-end of a page, i.e., interactions), Pandas (a library for handling different data structures in Python) and CSV (tool for dealing with files in CSV or comma-separated format). After this, the different static variables are listed: tournaments and years. Consequently, a loop is made for each tournament and each year, to modify the existing atptour.com URL (Uniform Resource Locator). Then, the web driver is started (an instance of Google Chrome) and the page is parsed looking for the table with the results. After it is found, the data is stored in a variable and then cleaned (eliminate unnecessary characters or text). Finally, the data is exported into a CSV file for analysis. The structure of the data is as follows: Year, Tournament, Ranking Winner, Winner, Result, Ranking Loser, Loser, Final Score and Comments.After the data is collected, the application of the algorithms commences. Firstly, the process begins in the “main” function. This function oversees all the code that is run in the Python script. It starts off with reading the CSV file that was created in the previous section of web scraping (historical tennis match data) and continues to “clean” the dataset (words to numbers). After the dataset is converted to numeric values, the feature columns and target columns are established. This is done with a simple loop that establishes every column except the result that is wished to be predicted as the features columns. On the other hand, it only establishes the result as the target column. With these columns defined, they can be passed on to the “print output” as parameters. A parameter is a value that is supplied to function when it is called. For example, if an "add" function was made, it could accept two numbers as parameters and return their sum. Consequently, the “print output” function begins with the output document “output.csv” being opened to clear any previous results the computer had calculated. It then opens the testing document, which contains the different results that need to be predicted. Consequently, it loops through these tests and creates an object for each type of linear regression model. An object is defined as a data structure that defines something. In other words, it is an instance of the data field. For example, if a student data type were created, each student would be an object of the data type (an instance). After each object is created, the objects are given the feature and target columns along with their corresponding values (fit function). With these values, the predictions are made with each linear regression model. These values are written to the corresponding document (output.csv). Additionally, with the keys and values document discussed in a previous section, the values are changed for a more user-friendly design. In other words, if the key for the value “Rafael Nadal” were 1, then the 1 is changed for “Rafael Nadal.” In the case of the ridge and lasso models, since the output is given in percentage, rules were written to give the output as a specific result. Which means that if the percentage of player 1 winning is greater than 49%, then it registers player 1 winning the match and vice versa.

Results and Discussion

For this project, particular test cases shall be applied to the project to see how accurate the prediction was. Matches that are not part of the historical data will be considered, that is, tournaments that occurred after the year 2022: Wimbledon, US Open, Roland Garros, and the Australian Open. After concluding the testing for the linear regression implementation, the evaluation of the test cases showed conclusive results as seen in Table 1:

Test #	Decision Tree	Lasso	Ridge
1	Player 1 Wins	Player 2 Wins	Player 1 Wins
2	Player 1 Wins	Player 2 Wins	Player 2 Wins
3	Player 2 Wins	Player 1 Wins	Player 1 Wins
4	Player 1 Wins	Player 2 Wins	Player 1 Wins
5	Player 2 Wins	Player 2 Wins	Player 2 Wins
6	Player 2 Wins	Player 1 Wins	Player 1 Wins
7	Player 1 Wins	Player 1 Wins	Player 1 Wins
8	Player 2 Wins	Player 2 Wins	Player 2 Wins
9	Player 1 Wins	Player 2 Wins	Player 2 Wins
10	Player 1 Wins	Player 1 Wins	Player 2 Wins
11	Player 1 Wins	Player 2 Wins	Player 1 Wins

Table 1
Prediction Results of Test Cases

After concluding the testing for the linear regression implementation, the evaluation of the test cases showed conclusive results. Table 2 demonstrates the accuracy of the results:

Model	Total Tests	Correct Tests	Accuracy %
Decision Tree	11	7	63%
Lasso	11	4	36%
Ridge	11	6	55%
Average	11	5.7	51%

Table 2
Conclusions of Test Cases

Table 2 shows that the algorithms were pretty accurate overall, with the average accuracy of a correct prediction being 51%. The best algorithm, decision tree, showed an impressive prediction rate of 7 out of the 11 test cases produced. On the other hand, the ridge algorithm demonstrated to be close in the race for the best overall prediction, with an accuracy of 55%. In other words, 6 out of the 11 test cases were correct. In sharp contrast, the last algorithm (lasso) showed a decline in prediction accuracy, with 36%, thus only predicting correctly 4 out of the 11 test cases. With these results, it can be confirmed that, with the implementation of these algorithms, a tennis match may be more accurately predicted. The sample test cases were produced with different players, tournaments, and rankings, demonstrating that the python script can adapt to different scenarios.

Conclusions

In conclusion, the project of web scraping and prediction with linear regression was an overall success, providing an overall accuracy of 51% and a useful tool for data gathering and collection. This project provides a useful tool for analysts, fans, and any person interested in the world of tennis, allowing the worlds of programming and sports to join forces and focus on producing more accurate predictions. In other words, this project represents the complete extinction of tennis upsets.

Future Work

Soon, many additions can be made to this project, specifically in two major areas: more features and other areas of implementation. Firstly, the more features collected via the web scraping algorithm, the more accurate the results will be. Columns such as surface of play, current tournament layout, age of players, climate, etc., can be added. Additionally, this project could be implemented in other areas in which prediction plays a crucial role. For example, in the world of firewalls and cybersecurity, these algorithms can be applied to historical malware attacks to help in the prevention of attacks based on the prediction they will occur, therefore allowing security staff to stop these attacks.

Acknowledgements

Firstly, I wish to express my gratitude to the different people who have helped me throughout the project. Without their help, in any type of form, this project could have been completed. These people include and are not limited to, my girlfriend, family members, friends, and many others along the way. Additionally, I want to express my utmost gratitude to Dr. Nelliud Torres for guiding me throughout this journey and bestowing upon me the tools necessary for the project. As well as the graduate school and the different members it composes.

References

[1] M. Schnur, “What size is a regulation tennis court?", Metro League,September 5, 2022. Available: https://www.metroleague.org/what-size-is-aregulation-tennis-court/

[2] History of Tennis, “The first official tennis match,” September 17, 2017. Available: https://thehistoryoftennisblog.wordpress.com/2017/09/21/introduction/

[3] R. Vora, “How to predict a tennis match?”, MatchStat June 29, 2023. Available: https://matchstat.com/predictions-tips/how-topredict-a-a-tennis-match/

[4] ATP Tour, “Homepage.” ATP Tour. Accessed Aug. 20, 2023. Available: https://www.atptour.com/

[5] IBM, “About linear regression.” Accessed Oct. 7, 2023. Available:https://www.ibm.com/topics/linearregression#:~:text=Resources-,What%20is%20linear%20regression%3F,is%20called%20the%20independent%20variable

[6] “Vector,” Encyclopædia Britannica, accessed Oct. 7, 2023. Available:https://www.britannica.com/science/vector-physics