Predicting Chicago Bulls Basketball Games Outcome

Joel Feddes  
joelfeddes@hotmail.com

***This project was designed to create a model that will predict the outcome of Chicago Bull’s basketball games by utilizing data science techniques such as web scraping and machine learning.***

# **introduction**

For this project, I have created a machine learning model to predict whether the Chicago Bulls will win or lose a game given the box score (statistics) of a game [1]. To complete this project, I utilized both web scraping and machine learning techniques.

# **related work**

An individual named Yash Gandhi, has completed an academic-based machine learning project used to classify the outcomes of soccer games in 2019. Although his project is different than mine, his project will help as a guide in how I structure my machine learning procedures [12].

# **problem description**

The purpose of this project was to accurately determine whether a Chicago Bulls game will result in a win or a loss. This is a useful project for individuals who are interested in predicting how well the Bulls will do in each season. This knowledge can be used by avid sports fans to help gauge their expectations of the Bulls upcoming seasonal performance. Also, it could help sports bettors place better bets.

# **methodology**

## **Python Web Scraper**

For data gathering, I created a Python web scraper [2]. This web scraper gathers data from seventy-two separate web pages [3]. Prior to performing web scraping, I verified the legality of web scraping from these web pages [4]. Each of these web pages contain various statistics surrounding the Chicago Bulls 2019-2020 regular season. However, my web scraper is designed to only scrape away the box scores from each of these seventy-two web pages. Once the box scores were scraped, I combined them into a single Pandas data frame.

## **Dataset Processing**

To refine my dataset, I performed data cleansing to include only the totals from each of the box scores. Also, I used the scraped data to compute calculated columns to provide myself with statistics such as “seconds played”. The box scores I scraped contain “minutes played” but does not contain the time played in seconds. As such, I will compute a column that calculates the number of seconds played. Thankfully, there were no null values present in the dataset. Next, I checked for, and evaluated, outliers to see if they needed to be removed or replaced from the dataset [5]. Finally, I was able to create a refined dataset to use for my machine learning models.

# **experiments**

To fine-tune the models, I performed experiments testing different parameter setting for each of the models. For Random Forest, I determined that 100 trees, 3 attributes at splits, and a class weight of balanced to do the best job classifying my dataset. For Logistic Regression, I used Lasso as the regularization type, balanced as the class weight, and set the strength of C to be 4. To help determine the outcome of a Chicago Bulls game [6], I fed my dataset into Random Forest [7] and Logistic Regression [8] machine learning algorithms. To refine Random Forest, I used 100 trees, and considered 3 attributes at splits. For Logistic Regression, I used Lasso as the regularization type. Also, I used a strength of 4 for C. Lastly, I set the class weights of both of my algorithms to balanced. Since the Bulls lost more games than they won in the dataset, this will help balance out this issue. The original attributes I fed to these models are SP, FGA, FG, FG%, 3P, 3PA, 3P%, FT, FTA, FT%, ORB, DRB, TRB, AST, STL, BLK, TOV, PF, and PTS [9]. After performing attribute correlation analysis, I removed TRB, FTA, 3P%, and FG% from the dataset [10]. To determine this, I used Pearson Correlations. Finally, to evaluate these models, I considered the accuracy score, F1 score, and confusion matrix [11].

Chart, waterfall chart

Description automatically generated

# **results**

For Logistic Regression, I was able to achieve a classification accuracy of about 75.5%, and an F1 score of 71.9%. Unfortunately, Random Forest did not do nearly as well of a job classifying the dataset. For Random Forest, I was able to obtain a classification accuracy of about 65.5% and an F1 score of about 54.6%.

# **conclusion**

When comparing the two models, it is plain to see that Logistic Regression did the better job out of my two models. Using Logistic Regression, we can now predict the outcome of Chicago Bulls basketball games based on a given box-score with about 75.5% accuracy.

Text

Description automatically generated with medium confidenceText

Description automatically generated

# **References**

|  |  |
| --- | --- |
| [1] | Today's Top Totals, "Basketball Prediction Algoithms and Sports Betting," 13 September 2019. [Online]. Available: https://todaystoptotals.com/basketball-prediction-algorithms-and-sports-betting. [Accessed 3 September 2021]. |
| [2] | S. Wu, "Web Scraping Basics," 15 July 2020. [Online]. Available: https://towardsdatascience.com/web-scraping-basics-82f8b5acd45c. [Accessed 15 September 2021]. |
| [3] | GeeksforGeeks, "How to Scrape Multiple Pages of a Website Using Python?," 3 March 2021. [Online]. Available: https://www.geeksforgeeks.org/how-to-scrape-multiple-pages-of-a-website-using-python/. [Accessed 2021 September 13 2021]. |
| [4] | Jami, "A Guide to Ethical Web Scraping," 2 December 2019. [Online]. Available: https://www.empiricaldata.org/dataladyblog/a-guide-to-ethical-web-scraping. [Accessed 4 September 2021]. |
| [5] | Humans of Data, "When Should You Delete Outliers from a Data Set?," 6 March 2018. [Online]. Available: https://humansofdata.atlan.com/2018/03/when-delete-outliers-dataset. [Accessed 20 September 2021]. |
| [6] | G. Nourse, "Machine Learning Algorithm for Predicting Outcomes of MLB Games," 6 May 2021. [Online]. Available: https://garretnourse.medium.com/a-machine-learning-algorithm-for-predicting-outcomes-of-mlb-games-fa17710f3c04. [Accessed 5 September 2021]. |
| [7] | A. Pandey, "Random Forest Algorithm in Machine Learning," 11 December 2020. [Online]. Available: https://adityapande107.medium.com/random-forest-algorithm-in-machine-learning-c4d9c3c585e1. [Accessed 15 September 2021]. |
| [8] | J. Brownlee, "Logistic Regression for Machine Learning," 1 April 2016. [Online]. Available: https://machinelearningmastery.com/logistic-regression-for-machine-learning/. [Accessed 15 September 2021]. |
| [9] | Breakthrough Basketball, "Basketball Statistics Definitions," 4 May 2021. [Online]. Available: https://www.breakthroughbasketball.com/stats/definitions.html. [Accessed 16 September 2021]. |
| [10] | C. Vaidya, "Which NBA Statistics Actually Translate to Wins?," Stadium, 15 July 2019. [Online]. Available: https://watchstadium.com/which-nba-statistics-actually-translate-to-wins-07-13-2019/. [Accessed 17 September 2021]. |
| [11] | A. Mishra, "Metrics to Evaluate your Machine Learning Algorithm," 24 Februrary 2018. [Online]. Available: https://towardsdatascience.com/metrics-to-evaluate-your-machine-learning-algorithm-f10ba6e38234. [Accessed 15 September 2021]. |
| [12] | Y. Gandhi, "Sports Predictor using Python in Machine Learning," 04 April 2020. [Online]. Available: https://www.codespeedy.com/sports-predictor-using-python-in-machine-learning/. [Accessed 28 September 2021]. |
| [13] | T. S. Figueira, "Web-scraping tables in Python using beautiful soup," 27 May 2021. [Online]. Available: https://medium.com/geekculture/web-scraping-tables-in-python-using-beautiful-soup-8bbc31c5803e. [Accessed 30 September 2021]. |