Premier Lacrosse League Analytics

Christopher Hsu

Denison University 2019, Data Analytics Capstone Project

Abstract:

The purpose of this research was to explore the topic of lacrosse analytics and to predict which team will win each individual game throughout the 2019 PLL (Premier Lacrosse League) season. I will be using game data from each of the 6 teams of the PLL separated by each week, which was taken from the PLL website. II used a random forests machine learning model for each week, using the data in each of the previous weeks to predict which team will win the game. With this method I am able to mimic prediction which team will win the game before it actually happened. Since the PLL is a brand new league I have no historical data. Since I have no historical data I will start my prediction model with week 2 and end with week 10. I created a model for each week, 2 through 10 with data from the weeks before that certain week. For example, week 4's model uses training data from weeks 1-3, and week 8 uses weeks 1-7 as training data. The overall performance of the model did pretty well with a 66.66% accuracy. The models highest accuracy for a week was 100% and the lowest was 33.33%. This predictive model would be able to be used at anytime over the course of the PLL season in the future, and will most likely produce better results as we collect more training or historical data. Overall, this model is better at predicting a game outcome than randomly guessing.

Introduction:

Sports analytics has been a very hot topic growing in popularity since the early 2000s. Sports analytics is a collection of relevant, historical, sports statistics when properly applied and used, it can provide a certain competitive advantage to a team or individual. Through the analyzation of the collected data, sports analytics can inform players, coaches and other staff to facilitate decision making both during and before sporting events. With advances in technology over the past years, data collection has become more in-depth and can be collected much easier than in the past. This lead to advancements in data collection which have allowed sports analytics to grow rapidly.

Sports analytics has been around for a while, but there is not much sports analytics being applied to lacrosse. With lacrosse growing in popularity around the United States and with sports better being expanded, there will be a higher demand for analytics providers. (Mann, 2019) I will be predicting how many wins each lacrosse team will have in the PLL, which is a professional lacrosse league that started in 2019 and just completed their first season. The main questions I will be looking into is, what are the main variables or factors apart of lacrosse that leads a team to more wins? Is a team's faceoff win percentages strongly correlated wins? Lacrosse players tend to assume that faceoff win percentage is important because it leads to more offensive possessions. Does the phrase “defense wins games” have any truth behind it where in sports if a team has an overwhelmingly strong defense, opposing offenses have trouble scoring. Or is the PLL just an offensive heavy league with a fast pace and lots of shots. These are just some of the possible questions I will be able to answer with my weekly models. I will be able to determine which factors play the largest role in a team winning a game by finding out which variable has the most impact on my model.

The introduction of the PLL has added a new dimension to the popularity of lacrosse. There are six PLL teams the Chaos, Whipsnakes, Archers, Redwoods, Atlas and Chrome. There are 10 weeks in a regular season so 30 total games.The PLL is becoming more popular and starting to become the primary professional lacrosse league. In the PLL there are no home or away teams. All the teams travel to different big cities or areas to play each other, so it is a neutral site. This lets us not having to worry about a team having home field advantage, or being comfortable in certain weather conditions. A team used to playing in California weather would not be as comfortable playing a team in Buffalo in the snow. This would lead to the buffalo team to have an advantage because they are used to playing in the cold weather and snow.

Domain Review

Sports analytics has been constantly evolving and are not just used for sports teams and players. Outside people or statisticians can use sports analytics to determine and predict who they believe will win a game and even a score of a match. This type of predictive analytics in sports lead to a significant impact on professional sports in relation to sports gambling or betting. This type of sports analytics has taken sports betting to new levels, from fantasy sports leagues or nightly/weekly wagers. These betters now have more information at their disposal to help aid decision making. Several companies and webpages have been developed to help provide fans with up to the minute information for their betting needs. Today there is an increased demand for sports analytics due to the increased popularity of sports media and information. Also, there are sports analytics providers that can provide data and analysis to let the user come to their conclusion. (Mann, 2019)

When looking at prediction models for sports, there is no exact perfect model for all because every sport is different in its own way. One of the most common models used for sports analytics is Neural Networks. Bunker and Thabtah use Artificial Neural Networks to predict sports results from NFL, AFL, Super Rugby, and English Premier League Football using data back to the year 2002. Due to the specific nature of match-related features to different sports, the results across different studies varied.

In a study by Constantinou, he looked at Dolores, a model designed to predict football match outcomes in one country by observing football matches in multiple other countries. This paper describes the model, which implements a rating system within the model. The rating system generates a rating score that captures the ability of a team relative to the residual teams within a particular league. The resulting ratings are then used as input to the model for match prediction. There are many different types of sports analytics models out in the world but Neural Networks is one of the more popular types.

Before creating a model we need to determine what are the variables and factors we want to use to create a model. With every sport, there are different types of specific variables you want to look at. With basketball, there are three-pointers, but there are not any in hockey. There are field goals in football but not in soccer. In professional lacrosse, there are 2 point goals. On the other hand, there are still variables that are universal in all sports. Leung and Joseph discuss in their paper what variables that were used and how they calculated them when predicting college football games. They used variables like win percentage, RPI (Ratings Percentage Index), which is a rating that takes into account their opponent's win percentage, expected probability, and updated probability after each game has been played throughout the season, and more all dealing with win-loss percentages. One limitation that was brought up when dealing with college football is the NCAA and players can only play for four years. Generally, a good player in a good football program is likely not able to get significant playing time until his third or fourth year in college, further limiting the amount of relevant data. Also with freshman or first years on a team, we do not know the impact they will have. One workaround would be to only look at very recent data spanning from one to three years. In this study’s case we will be looking at the most recent PLL season, 2019, so this would include the younger generation of players, who are less injury-prone. There was a smaller percentage of older players that joined the PLL.

In another article, by Cokins and Schrader, they looked at Improved predictive models for play tactics and how to measure them. Cokins and Schrader give many possible variables for building a next-play model in football like, what yard line is the ball on and which hash mark? What are the down number and yard distance to a first down? What is the score differential at the time of this play? How much time is left? What is the offense formation? What is the defense players’ formation? What is the outcome of each play? The sequence of plays can also be relevant. Another possible variable they talk about is with “explosive plays” where the offense gains 16+ yards passing or 12+ yards rushing in Football. This is a variable that applies to both offense and defense. With offense, there is explosive plays percentage and we could also look at attempts. On defense, there are explosive plays given up percentages.

When looking at these different variables come up with certain variables that can measure certain situations and aspects of the game of lacrosse. We can use the universal variables to measure a team's record like wins, loses, win/loss percentages, points per game, points against per game. We could also create a rating system that could possibly measure a team's performance in a certain non-quantifiable aspect of the game. We could give each team a ball-movement rating by recording how many assisted goals there are. Ball movement is very important to lacrosse and scoring because, with quick ball movement, it tends to get the defense out of position leading to goals. Also, implementing the play tactics variables like explosive plays. We could measure things like momentum with scoring within a minute of each other and points off turnovers. We could also measure tempo or speed by recording fast-break goals within a minute of a face-off. We can measure the specialist part of the lacrosse like face-off win percentage, man-up conversion percentage, man-down defense percentage, and average amount of penalties from each team. Then we get into individual stats like one point goals, two-point goals, caused turnovers, save percentage, clearing percentage. But there are still some unknowns when looking at the Premier lacrosse league, like how will the newly drafted players perform and their impact on the team? Also as players get older, are they going to still have the same impact as when they were younger? These are some of the questions we have to consider before creating a model.

Data

The data I am using is team data from each game from the PLL separated by week. I web scraped the data from the PLL online public websites. I am taking the team data from this past PLL season because there is no historical data. Since the PLL had just concluded its first season I will not have enough data to create a model determining the overall season. Since I do not have enough data, I will use the PLL data from each week as training data. I have multiple different datasets. The first one is a dataset holding all the teams overall season stats. The second dataset is of gamestats, that has all the stats for each team from each game. The last one is the matchup stats, which has data on each matchup and the stats the team has against certain opponents. The variables I am have collected for datasets are team, wins, losses, total points, total 2pt goals, total assists, total shots, shot percentage, ground balls, turnovers, caused turnovers, save percentage, points against, emo (extra man offense) goal percentage and faceoff win percentage. After I collected all the data, I calculated the per-game averages for points, 1pt goals, 2pt goals, assists, shots, ground balls, turnovers, caused turnovers, and points against for each team.

Statistical Methods

Random Forests

I will also be using Random Forests to determine how many wins each team will get. Random forests method in sports analytics typically applied to game outcomes. While conducting my domain review I realized that the two most popular models for sports analytics have been Neural Networks and Random Forests. While creating both of those models there was a clear difference that Random Forests created a better model. This may be due to the lack of data because Neural Networks tends to create a more accurate model with larger data sizes. Random forests are the aggregation of a large number of classification or regression trees. This method can be used both for classification, determining which team would win or lose, and regression, score predictions, purposes. The trees are grown independently from each other, and the end model, predictions of the individual trees are aggregated. (Schauberger & Groll, 2018)

The Random Forests input variable will be the teams' statistics playing each other and the output is the team that is predicted to win.

Model Validation

After creating the models I will need to validate them to see how well the model did. I have all the scores on outcomes of the PLL season from the PLL website, so the first validation would be how accurate the model was for predicting the wins of each week. I created a confusion matrix to examine the overall accuracy, sensitivity, specificity, and Kappa score, how much better is the model than randomly guessing.

Limitations

There are many limitations to my model. The first limitation is with the amount of data I had. Since the PLL just had their first season there is not enough data on each PLL team. The second is with the type of data recorded in lacrosse. There are many variables that are just not recorded like time of possession, clearing percentage, Man down defense percentage. Those are just a few important variables that I would have liked to have used when creating my models.

If I had data on time of possession I could have also measure tempo or pace by recording fast-break goals within a minute of a face-off. Or measure pace like in Basketball by using Team time of possession vs opponent time of possession. I could have measured the specialist part of the lacrosse like man-up conversion percentage, man-down defense percentage, and the average amount of penalties from each team. I also did not look into individual player stats like one point goals, two-point goals, assists, caused turnovers, and save percentage. But there are still some unknowns when looking at the Premier lacrosse league, like how will the newly drafted players perform and their impact on the team? Also as players get older, are they going to still have the same impact as when they were younger?

There are also some different types of data that I would have liked to see. In another study, Akiyama, Sasaki, and Mashiko looked at GPS and heart rate data on the Japanese lacrosse team. They looked at data like total distance moved by each player, their run times, sprint time (<21.6 km/h), run time (14.4–21.59 km/h), jogging time (7.2–14.39 km/h), stand/walk time (0–7.19 km/h), heart rate, and acceleration. With these types of GPS variables, they were able to look at each players work to rest ratio which reflects total high/moderate running speed exercise versus total low-speed running time. With all of these different types of variables from the studies above, it can all be applied to lacrosse analytics. I would be able to see which individual player is the superstar for each team and how well they play against certain teams if they are tired. Or even just look at different irregularities when the certain player goes against different matchups from each team.

Ethical Statement and Protocol

Data Gathering Procedure

The data I will be using is team data from the PLL. I web scraped the team data from the PLL online public websites. I am taking the team data from this past 2019 PLL season. I do not need IRB approval because I am looking at professional lacrosse team stats and not an individual player. Plus this is all public information, freely available that anyone can find online.

Data Storage Procedure

Since I put a lot of time and effort into web scraping and organizing the data I collected, data security is important. I will not be publishing this data online freely, and I will be keeping this data for my personal use only. I will be storing this data on my computer and personal drive.

Reporting of results

With the reporting of results, I will be predicting how many wins each PLL will have at the end of the season. These results will only be using the team stat variables such as points, 2pt goals, assists, shots, ground balls, turnovers, caused turnovers, and points against. I will not be publicly publishing my model due to how my models could become publicly used in sports betting or gambling scenarios. Also, if these professional sports teams see that in my model, if they had the lowest scores, the players could lose motivation, leading them to get released.

Results:

Preliminary Findings:

For my early results, I decided to look at the correlation between all the variables from the PLL teams this past 2019 season. I decided to create a correlation heat map and look at PCA with all the variables (Appendix 1).

PLL Correlation Heat Graph

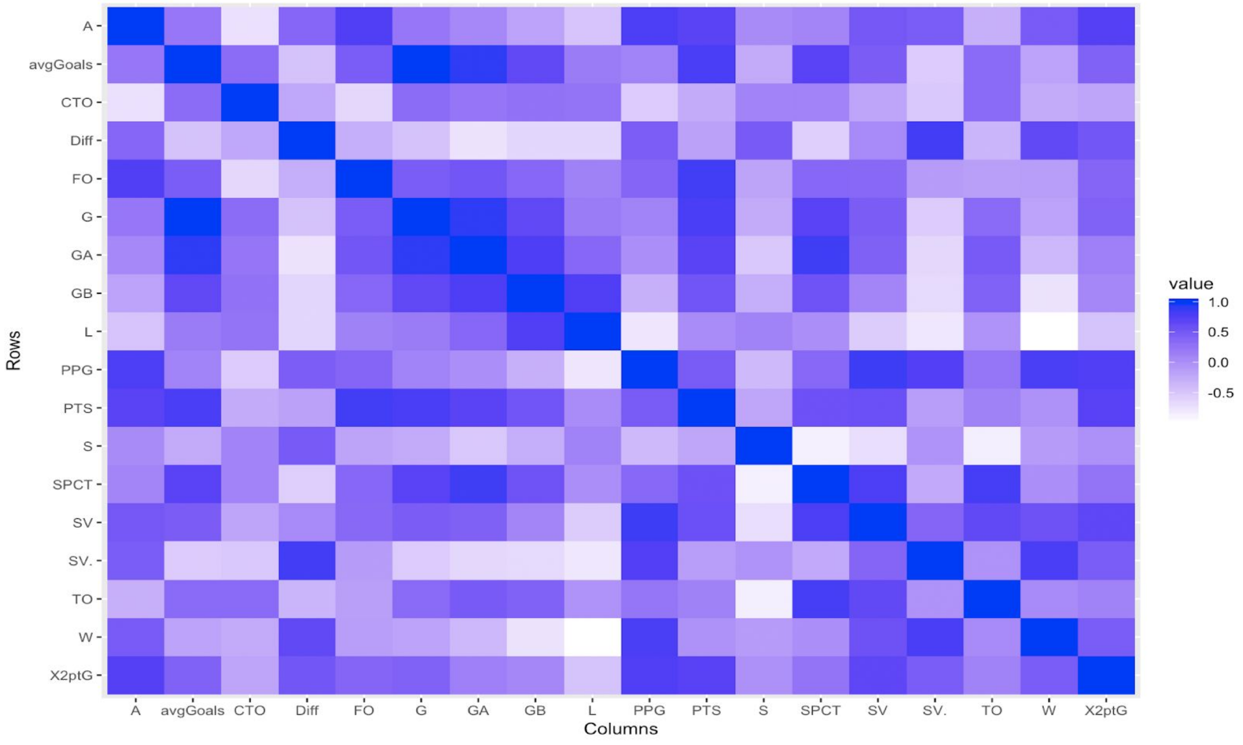


Figure 1.

Since I am mainly focused on wins, I wanted to see which variables are highly correlated. The first thing I notice is that Goals is very slightly negatively correlated with wins, which is not what I was expecting. But assists are pretty positively correlated with winning, which leads me to believe that these winning teams have good ball movement. The next variable I notice is ground balls are very negatively correlated. This could be due to teams who get more ground balls are not converting those possessions into points or maybe have a lot of errors and collect their ground balls. Saves and Save% are very highly positively correlated which means the good teams have good goalies who sometimes make saves when they are not expected to. 2pt goals are highly correlated with wins which means the top teams have better shooters and are converting on their 2pt attempts.

Since Goals and wins are not very correlated I decided to make a graph of wins and average goals per game in the PLL.

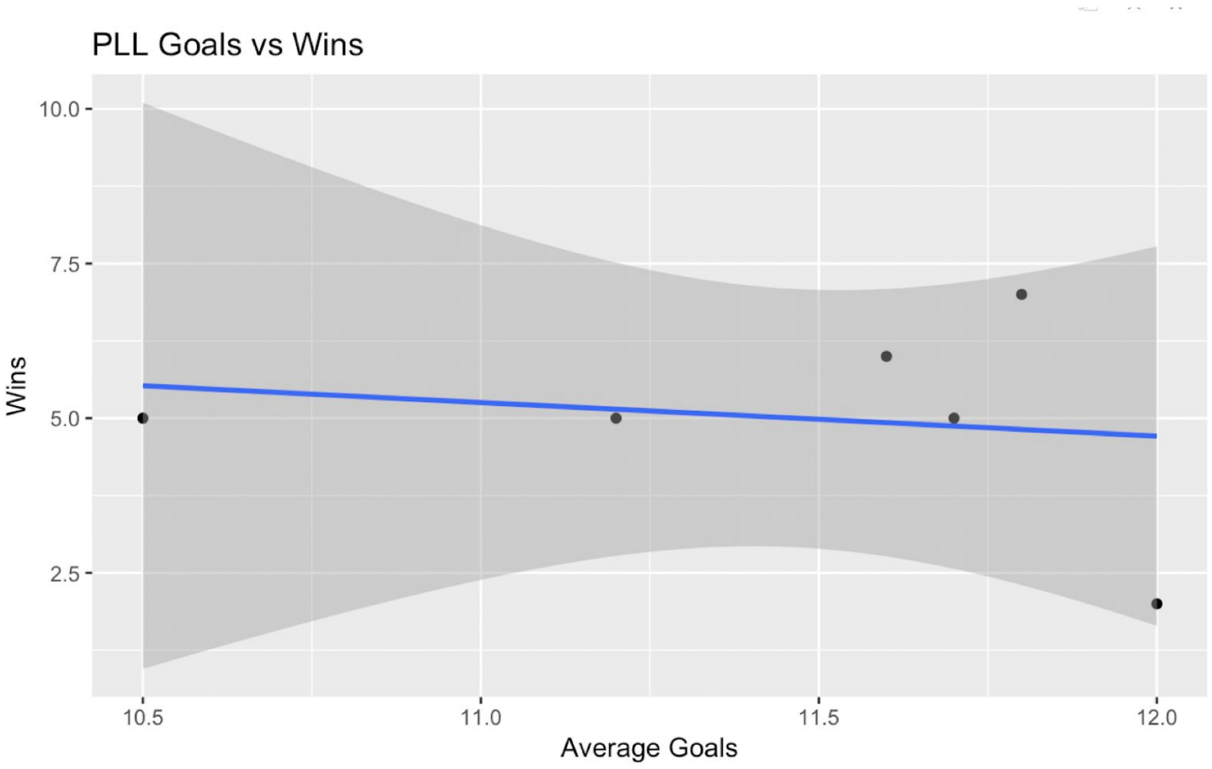


Figure 2.

If we took out the team with only 2 wins who are averaging 12 points a game, we would believe goals are very highly correlated with wins. The team with only 2 wins have been in every game and is a high scoring team, but lost most of the time. The team with 5 wins and are averaging 10.5 points a game must be a more defensive team because they would win games and not have to score as many goals as the other teams.

Prediction Models

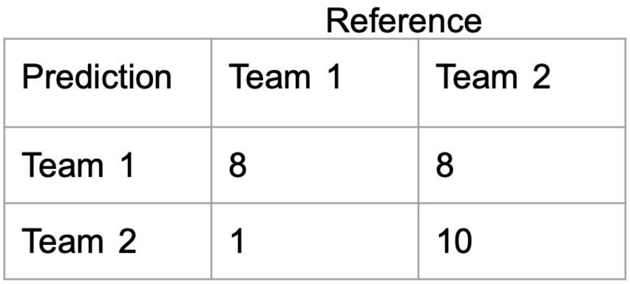
Random Forests

For my model I created a Random Forests prediction model using the game stats from each week. I used team stats variables representing both teams in my model and also included how many wins each team had before the game. There was a different model created for each week of the season starting at week 2. So as the season progressed the model was trained with more data. I created a table to show the models accuracy of each week.

Week	RF Model Accuracy
2	66%
3	100%
4	66%
5	33%
6	66%
7	66%
8	33%
9	66%
10	100%

Figure 3.

In Figure 3. For week 2 the model was 66% accurate which means it predicted 2 out of the 3 names correctly. We assume the model will get better over time as it collects more game data. In week we can see that our accuracy dropped to 33%, that is partially due to an upset happening during the season where the lowest ranked team that only won one game, won their second game against a higher ranked team.



Reference

Prediction	Team 1	Team 2
Team 1	8	8
Team 2	1	10

Figure 4.

I created a confusion matrix in Figure 4. Based on the overall model's prediction. The model had an accuracy of 66.66%. It had a sensitivity of 89% or true positive rate, where our model said Team 2 will win, when Team 2 actually won. It got a specificity of 56% or true negative rate, where our model said Team 1 will win, when Team 1 actually won. It got a Kappa of 37%, which is pretty good because it is saying our model is 37% better than just randomly guessing which team will win.

Importance Variable Plot

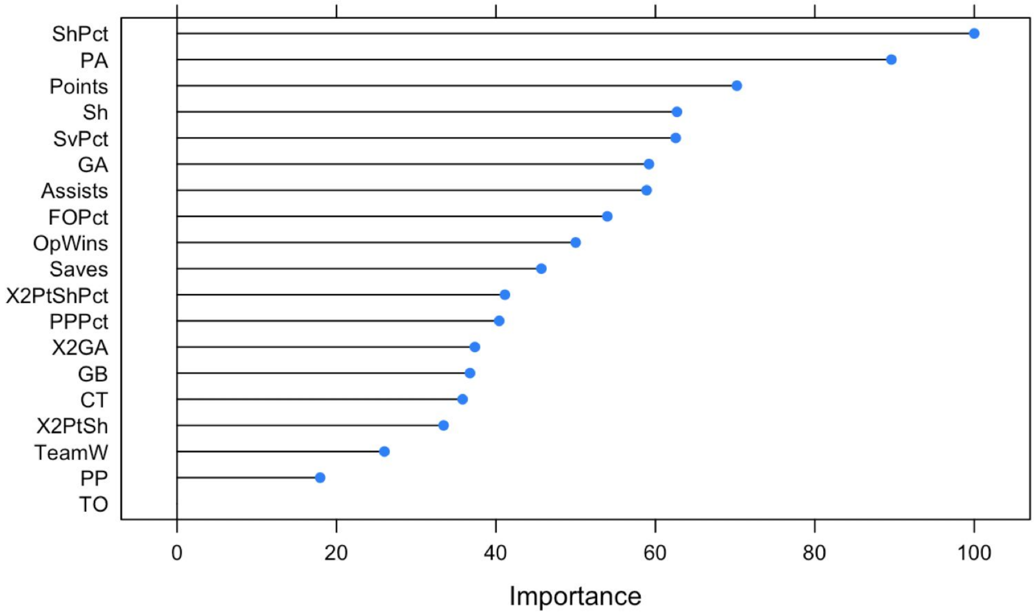


Figure 5.

I also created an importance plot for the final Random Forests model of week 10, Figure 5. This model shows the weighted importance each variable has on a certain team winning a game. We can see that Shot Percentage and Points Against are the two most important variables in our model. For Shot Percentage we can assume that the winning teams are taking higher quality, higher percentage shots. These teams are typically taking shots that are more likely to score than others. Typically a high percentage shot is a shot within 10 yards of the goal, and a low percentage shot is over 15 yards away. Higher percentage shots are typically assisted goal where a player is open and has enough space to shoot. We can see that assists are in the upper half of important variables. With Points Against, we can assume the winning teams have a good defense, not allowing opposing teams to have high percentage shots. This would lead to having less goals scored against making it easier for that team to win. I did expect Face-Off percentage to be higher because typically the more face-offs a team wins, means more possessions, and that means there are more opportunities for that team to score.

CONCLUSION:

Overall, the two models did better than I expected. Having a Kappa at 37% is pretty good and it is significantly better than randomly guessing. I do believe that over time, as the PLL completes more seasons, this model will become better. Over time, the PLL may collect more in dept statistics on each team, which would benefit my model and lacrosse analytics.

Also, another problem with sports analytics with continuous sports like lacrosse is there are many non-quantifiable variables and factors that go into the game. In sports there are many factors that you cannot quantify like momentum, coaching, and even sometimes luck. A stop-start sport like baseball is very analytics heavy because it is easier to get a more accurate representation of a player with his individual statistics and matchups. The stop-start sports tend to minimize these non quantifiable factors involved in a game.

REFERENCES:

McCabe, A., & Trevathan, J. (2008, April). Artificial intelligence in sports prediction. In Fifth International Conference on Information Technology: New Generations (2008) (pp. 1194-1197). IEEE. https://www.researchgate.net/profile/Jarrod_Trevathan/publication/220841301_Artificial_Intelligence_in_Sports_Prediction/links/00b7d5154fe649278f000000.pdf

Schauberger, G., & Groll, A. (2018). Predicting matches in international football tournaments with random forests. Statistical Modelling, 18(5–6), 460–482. https://doi.org/10.1177/1471082X18799934

Mann, R. (2019, March 31). The Marriage of Sports Betting, Analytics And Novice Bettors. Retrieved from https://sportshandle.com/sports-betting-in-us-data-analytics-industry/

Sports Data Mining: Predicting Results for the College Football Games. (2014, September 13). Retrieved from https://www.sciencedirect.com/science/article/pii/S1877050914011181

Cokins, G., & Schrader, D. (n.d.). The Sports Analytics Explosion. Retrieved from https://www.informs.org/ORMS-Today/Public-Articles/February-Volume-44-Number-1/The-Sports-Analytics-Explosion

Akiyama, K., Sasaki, T., & Mashiko, M. (2019, June 1). Elite Male Lacrosse Players' Match Activity Profile. Retrieved from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6543992/

Bunker, R. P., & Thabtahb, F. (2017, September 19). A machine learning framework for sport result prediction. Retrieved from https://www.sciencedirect.com/science/article/pii/S2210832717301485

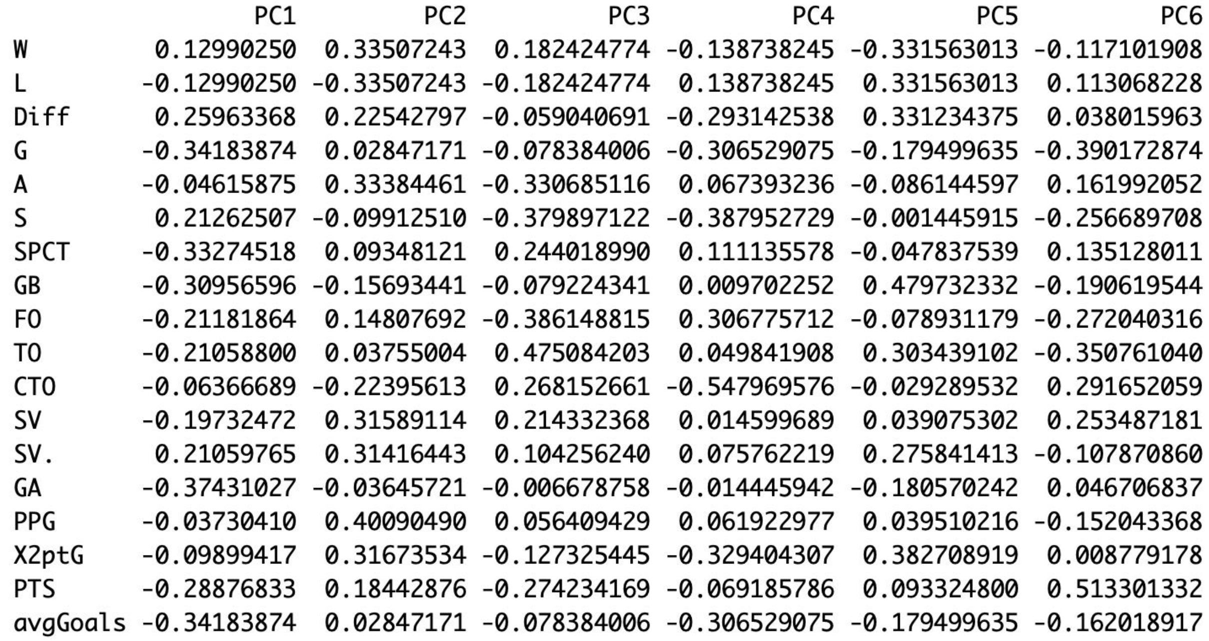
Constantinou, A. C. (2018, May 3). Dolores: a model that predicts football match outcomes from all over the world. Retrieved from https://link.springer.com/article/10.1007/s10994-018-5703-7

Data websites

PLL: https://stats.premierlacrosseleague.com

APPENDIX:

A.1: PCA Table of PLL Team Stats



```
                PC1         PC2          PC3          PC4          PC5          PC6
W        0.12990250  0.33507243  0.182424774 -0.138738245 -0.331563013 -0.117101908
L       -0.12990250 -0.33507243 -0.182424774  0.138738245  0.331563013  0.113068228
Diff     0.25963368  0.22542797 -0.059040691 -0.293142538  0.331234375  0.038015963
G       -0.34183874  0.02847171 -0.078384006 -0.306529075 -0.179499635 -0.390172874
A       -0.04615875  0.33384461 -0.330685116  0.067393236 -0.086144597  0.161992052
S        0.21262507 -0.09912510 -0.379897122 -0.387952729 -0.001445915 -0.256689708
SPCT    -0.33274518  0.09348121  0.244018990  0.111135578 -0.047837539  0.135128011
GB      -0.30956596 -0.15693441 -0.079224341  0.009702252  0.479732332 -0.190619544
FO      -0.21181864  0.14807692 -0.386148815  0.306775712 -0.078931179 -0.272040316
TO      -0.21058800  0.03755004  0.475084203  0.049841908  0.303439102 -0.350761040
CTO     -0.06366689 -0.22395613  0.268152661 -0.547969576 -0.029289532  0.291652059
SV      -0.19732472  0.31589114  0.214332368  0.014599689  0.039075302  0.253487181
SV.      0.21059765  0.31416443  0.104256240  0.075762219  0.275841413 -0.107870860
GA      -0.37431027 -0.03645721 -0.006678758 -0.014445942 -0.180570242  0.046706837
PPG     -0.03730410  0.40090490  0.056409429  0.061922977  0.039510216 -0.152043368
X2ptG   -0.09899417  0.31673534 -0.127325445 -0.329404307  0.382708919  0.008779178
PTS     -0.28876833  0.18442876 -0.274234169 -0.069185786  0.093324800  0.513301332
avgGoals -0.34183874  0.02847171 -0.078384006 -0.306529075 -0.179499635 -0.162018917
```