Data Science in Duke Athletics Performance

Business Understanding: Duke Athletics has invested heavily in equipment that athletes can use to both train and capture data around performance; this includes specialized weightlifting platforms, monitors to track speed and agility while performing open-field drills, and camera equipment to capture basic telemetry data while the athletes are practicing their sport. While Duke has been enabled in capturing this data, the coaching and training staff is not sure what to do with all of their data. Furthermore, the coaching staff does not store this data in any meaningful manner, opting simply to capture the data in locally-stored spreadsheets with some similarities, but without standardization, from year to year. Exacerbating the problem, individuals who fill the role of "Data Scientist" for each sport are siloed and arbitrarily distributed; some higher-profile sports (e.g. Men's Basketball, Football) have dedicated teams and individuals, while other "Olympic"[1] sports have one individual shared across multiple teams. Some teams (Lacrosse, Rowing, Track and Field) store a subset of their data available, but leverage only the most arcane plots and charts to understand and interpret the data. Other sports forego data collection entirely, coaching instead by "feel".

Our group worked closely with several staff members at Duke Athletics who work on Sports Performance across these various teams. Duke Athletics leadership is aware that their teams are under-utilizing data and suspects (correctly) that even modest investments of time, energy, and resources into this space could yield easily-actionable and implementable insights. As a nonprofit undergraduate program, Duke Athletics is also hesitant to invest heavily in an external consulting firm to do the initial work of cleaning data and proposing future-case best

[1] "Olympic" sports is the designation given to every sport besides Men's Basketball, Women's Basketball, and Football.

practices, preferring instead to leverage more accessible resources to deal with structural elements such that a subsequent (and future hypothetical) Data Science professional services firm might instead dedicate their time exclusively to modelling and insights. Duke Athletics is also eager to see an example case study of how existing data can be standardized, joined with other readily-accessible datasets, and converted into an actionable insight; the hope is that this can be used as a selling point on the value of this type of work. Given the aforementioned situation, our group's final project for DEC618 is a natural fit to serve those needs.

We engaged with Carl Christensen, Assistant Director for Sports Performance for Olympic Sports, earlier this quarter; together, we identified Duke Lacrosse as the ideal partner for this engagement. The team has a uniquely high ratio of data collected to insights gleaned; that is, they have access to substantial time-series performance analytics data and robust publicly-available game and season-specific data, but have limited storage methods and have not been able to glean any actionable insights to date. Extremely basic plots have helped identify trends, but these have largely been less helpful (e.g. higher lifting is correlated to higher body weight). Our team worked with Duke Athletics to gain an understanding of the data available and the types of analyses the coaching staff is looking for. The dataset we were provided by the Men's Lacrosse team captured strength data across 13 variables, 97 players and on-field performance across seven seasons (from 2013-2019)[2]. The staff would like to better understand how certain strength and speed variables can predict strength progression in the weight room and would also like to identify strength exercises better predict on-field performance, drawing parallels in on-field performance and off-field performance. Our research and modeling aims to answer both questions.

[2] This period is particularly insightful as Duke Lacrosse was extremely successful on the field during this period, including two National Championships (2013, 2014) and a second-place finish (2018). This also means the data includes more games, due to the long postseason runs by the team.

On a more macro level, Carl's team has also asked that we leverage lessons learned from this class to better inform how Duke Lacrosse (and, by extension, the rest of Duke's teams) can better structure, store, and use data going forward. Fundamentally, the staff is quick to acknowledge that they are not true Data Scientists, and they hope that our project and recommendations can give insights into the types of data that is collected, the manner in which it is to be stored, and how these actions can better position their future-state teams to deliver higher-quality results on the field. In that way, the aforementioned (and hypothetical) future MBA and/or consulting team would be best positioned to efficiently target those questions and answers and would not invest effort in simply cleaning and preparing data. The insights of our project would also illustratively demonstrate and exemplify the value of those types of investments, allowing Carl and his team to sell the investment those types of high-end data analytics projects projects might entail and make those costs more palatable to administrators.

Data Understanding: Our data was sourced from two sources, primarily:

- **Strength and speed performance data**: Date-specific strength and speed data was provided by Carl and the Duke Athletics Sports Performance team. The data is captured each year during training sessions in both August and November where trainers capture strength (squat, bench, clean lifts, various functional exercises) and speed/agility (shuttle run "5-10-10") data. Raw data was provided to the team and then converted into structured data for R. A glossary of this data is available in **Appendix I**. **Appendix II** shows an example of this data in its raw form, and highlights some of the initial challenges we ran into when cleaning the data.

- **On-field performance YoY data:** this data was scraped from online datasets that show full-season, on-field statistics (all variables defined in Variable Glossary, attached). This data is typically collected each game by a statistician, compiled by the home-team's Sports Information Director, and then shared publicly. Duke Lacrosse does not collect game-specific data beyond this information.[3] Sourcing information can be found in **Appendix IV**.

Data Preparation: The Lacrosse dataset was robust and large enough that we were excited to start finding conclusions, though unfortunately the data was not formatted in a usable manner when furnished by the team. Specifics of the initial format can be found in **Appendix II**; notably, there was a separate file for each year, with poorly-organized individual data from player to player. This meant that a player's freshman year would feature one year's worth of data, but the subsequent year would feature the same data in a different file - not allowing for good data year-over-year. Given that there were fewer than 100 total players (once multi-year duplicates were removed), we decided to manually clean the data by selecting only the *last* year that each player appeared (as this would also have all historical data available for that player from all prior years) and manually removing them from each prior year. Substantial effort was also spent normalizing variables that had names changed year-to-year,[4] with notes made as a best practice to standardize these numbers in the future. Some variables were only collected in a subset of the years; in most cases, we dropped these variables due to insufficient data. Finally, headers were repeated regularly throughout to aid with scrolling while viewing the data; these rows were dropped.[5] While in this case, we were able to (albeit tediously) solve many of these issues, we identified them as best practices that the team should engage in future years,

[3] Or, if that data *is* collected, it was not communicated or made available to us.
[4] E.g., "Pound for Pound" one year vs. "Pound 4 Pound" in the next year
[5] Humorously, this issue could probably have been avoiding using Freeze Panes function in Excel

as this is not a sustainable practice. Lastly, to ensure privacy compliance, we removed all player names and replaced them with ID numbers for tracking purposes.

After talking to Carl and his team, and gaining a better understanding of the types of metrics that the coaching staff was looking for, we created three calculated variables that would aid in identifying actionable KPI's:

- **Turnover Efficiency:** The number of caused turnovers divided by the number of turnovers committed, for a given player
- **Start Percentage:** The percentage of games played that that player started in
- **Points per Game:** Sum of Goals and Assists, divided by the total number of games in which that player played

Given that much of the data had been stored in local Excel files, and in an effort to expedite the process (despite the natural labor intensity this approach entailed), we leveraged Microsoft Power BI to finalize the dataset into a format in which we were comfortable importing it into R. After this process was complete, our dataset featured a time-series history for each player, with measured strength data, on-field performance data, and calculated variables. We were now ready to pivot to R to perform more in-depth mining and analyses on this data.

Pre-Modeling Edits: Out of Sample size included 41 of the total 273 rows. This represents 15% of the population. We would normally like to use 10% of the data as a holdout, but our team concluded that this would be too minimal for the model. It is prudent to assume that as data collection strategies improve this might be able to be reduced to 10% in a future-state. We had 230 NA's out of 273 rows (player-years) for speed data because the sports performance team just started collecting speed data in 2018. We chose to drop all speed variables as they would only play a role for players in the 2019 season.

There was a “null” value for every player-year in our data. We addressed this by writing script to replace “NA’s” for November with August data, and vice-versa, so long as the data resides within the same calendar year. We did not pull data across years, merely months from August up to November or from November back to August.

Modeling: We first looked at intra- and inter-dataset correlations to find any initial interesting trends; these correlation matrices can be found in **Appendices V, VI, and VII**. **Appendix VIII** shows correlation matrices across the calculated variables, again looking for initial interesting trends. We decided that we wanted to predict points (goals + assists) per game for players based on the data that Duke Athletics has collected over the last few years. We joined this with on-field performance data and created a points per game metric. In order to choose a model for our prediction, we ran a generalized linear regression, lasso, and post-lasso. We chose a linear regression because we do not have any binary variables to predict. One con of using a linear regression is the number of variables we have and how highly correlated many of them are with each other.

We created the models using training data that was a subset of our overall data. After creating these models, we tested their out-of-sample performance by looking at the mean difference in the prediction and the actual for our holdout data. A plot of the OOS performance results can be found in **Appendix XI.** As predicted based on our correlation, a subset of the variables found using lasso performs better than our model with all variables.

Given that lasso showed the best OOS performance compared to the null model, we decided to continue with a cross-validated lasso (lassoCV) and predict the points per game for a player given the subset of attributes lasso dictated should be kept in the model. A fitting graph of the lassoCV can be found in **Appendix XII**.

We created lassomin by using the lambda.min and our test/holdout data. We then ran a prediction using lassomin and found predictions ranging from 0.008 to 11.9 points per game with a mean of 1.54. Looking at a histogram (**Appendix IX**), shows that there are outliers at the top end with most players having 4 points or fewer per game.

After running lasso, we decided to run principal component analysis (PCA) to see what variables, when uncorrelated, impact points per game. The results of the PCA show that there are two main sets of factors that explain the variance, as shown on the graph in **Appendix X**. The first, and largest, factor set is all about the weight lifting metrics for the player (total lift/squat/bench). The second factor set is a subset of on-field performance metrics, which can only be used for prediction for players that have statistics from last season (e.g. not freshmen). Due to this limitation, the weight lifting measures should be the main focus for Duke Athletics.

Evaluation: The OOS performance of our lasso model is 99.74%. This is very high, but we recognize that our sample size is small and the models should trained and tested again after data are collected for more players. Even with this caveat, a business case can be made to use the model to predict which players they can expect will have more points per game if they get playing time, and this should lead to more wins for Duke Lacrosse than if they did not consider each player’s attributes and how they can predict the expected points per game. Since Duke Athletics is in the business of winning sporting events, we do not have a monetary ROI calculation for them in evaluating if they should deploy this model or not. Instead, we feel the value proposition is focused on athlete wellness and properly setting goals and expectations that are achievable.

Given the number of NA's, we need to have a predictive framework prior to them coming into school if they do not come for an August workout or were injured at a certain point.

Business Recommendations: Beyond our findings that are illustrated above, we have a few primary recommendations for the coaching staff, and by extension Duke Athletics administration, to consider, when viewing the data available through a more macro lens. We believe that these should be applied as best practices across the athletic department, which can also help with standardizing and centralizing data storage in migrating the department away from offline and local spreadsheets.

For lacrosse specifically, points per game predictions based on strength numbers (recorded in the fall) will allow coaches to make more informed decisions in a season's early games when on-field performance data is very limited. Not only will this allow coaches to make better decisions around who starts but it also allows for a better allocation of playing time.

Data Availability: It is clear that the team is not collecting or storing data in a manner that is actionable; while our previously-described data cleaning was able to create a more usable dataset for the limited timeframe on which we were working, this process is neither sustainable nor repeatable. The data that is collected should be transformed out of excel and into a more versatile data system, with a single record for a given player can be updated each session. Once this is in place, it would make sense to also increase the frequency of data collection to provide more precise time-series for future models. Once this is in place, reports can be pulled that closely align to the existing spreadsheets without disturbing the underlying dataset.

Additional Variables for Consideration: It seems that the coaching staff's primary concerns are to maximize lift gains, as the KPI's that they track are only those most basic numbers. Indeed, the only calculated KPI is the “Pound for Pound” metric; this is not an actionable insight, as it assumes a linear relationship between body weight, which is unlikely to be true. It would also be prudent to collect more robust in-game data, either internally or by scraping publicly-available live game feeds. Variable naming conventions should also be established and adhered to, among and across all sports. Finally, we suggest that minutes played be tracked more closely; our finding that Points per Game is highly correlated with turnovers and players that play more games also score more points seems counter-intuitive; if there was a way to quantify the amount that a player played in a game beyond a binary “Yes they played” vs. “No they did not play”, this could be better resolved.

As a concluding sentiment, it’s important to note our belief that **the variables Duke Lacrosse is collecting now are not the correct type of data to drive actionable insights from a causal data study**[6]. Our understanding of the data that exists for other teams is that this problem persists; to better inform decision-making on the field and more effectively leverage the “smart” equipment with which their teams have been enabled off, coaches should track data differently.

[6] For example - the coaching staff is concerned with how many turnovers are caused vs. turnovers committed, but often those who cause turnovers are in different positions than those who are in a position to commit a turnover.

Appendix I: Variables Glossary:

Strength Performance (all stats represent testing numbers at one point in time)

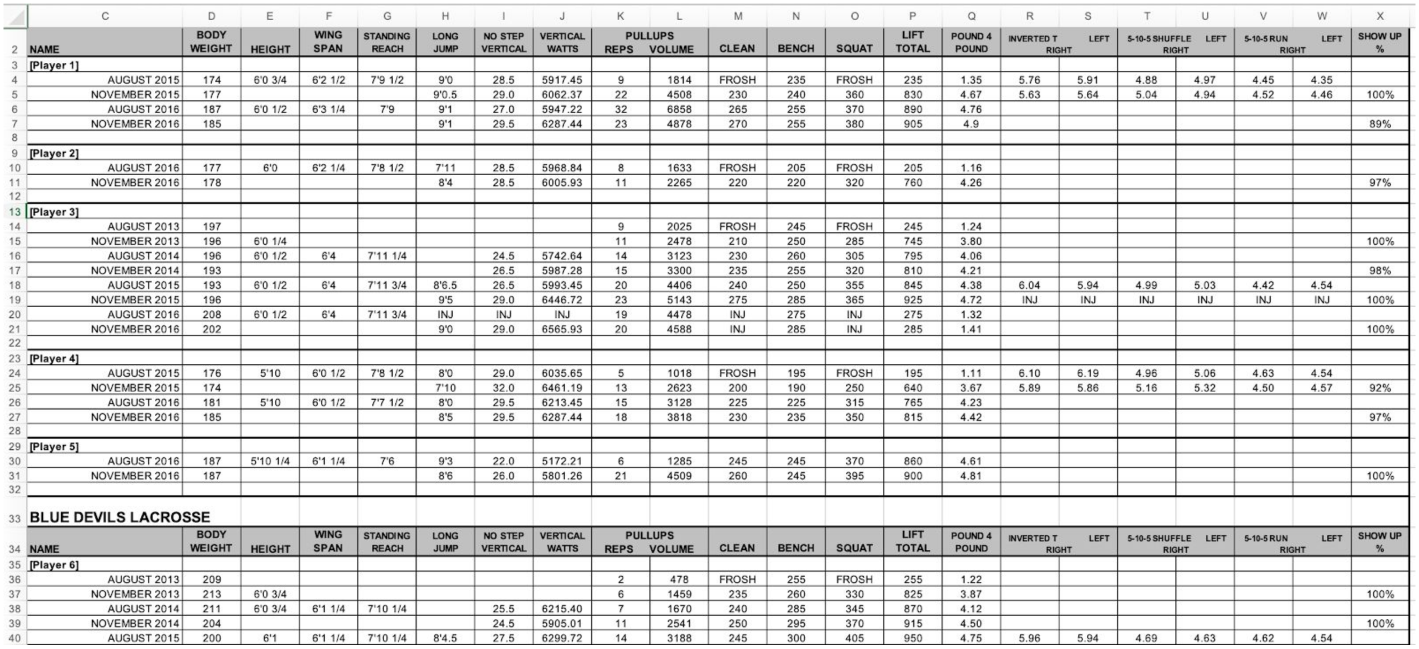
- **BODY WEIGHT** - Weight of the player, in United States pounds
- **NO STEP VERTICAL** - Vertical jump height without a running step
- **VERTICAL WATTS** - Vertical jump divided by weight
- **CLEAN** - Clean jerk lift, in US pounds
- **BENCH** - Bench press, in US pounds
- **LIFT TOTAL** - Sum of all lifts, in US pounds
- **POUND 4 POUND** - total number of pounds lifted divided by the player's weight
- **SQUAT** - Standard squat, in US pounds
- **5-10-10 - LEFT** - Agility Test. Variable dropped due to insufficient data
- **5-10-10 - RIGHT** - Agility Test. Variable dropped due to insufficient data
- **SHOW UP %** - Percentage of times that the player showed up to mandatory practices. Not consistently tracked, Variable dropped due to insufficient data
- **PULLUPS - REPS** - Number of pullups the player can do. Variable dropped due to insufficient data
- **PULLUPS - VOLUME** - Total number of pullups the player can do. Variable dropped due to insufficient data

On-field Performance (all stats represent a given season)

- **Games Played** - Number of games in which the player made an appearance
- **Games Started** - Number of games in which the player started
- **Goals Scored** - Number of goals scored

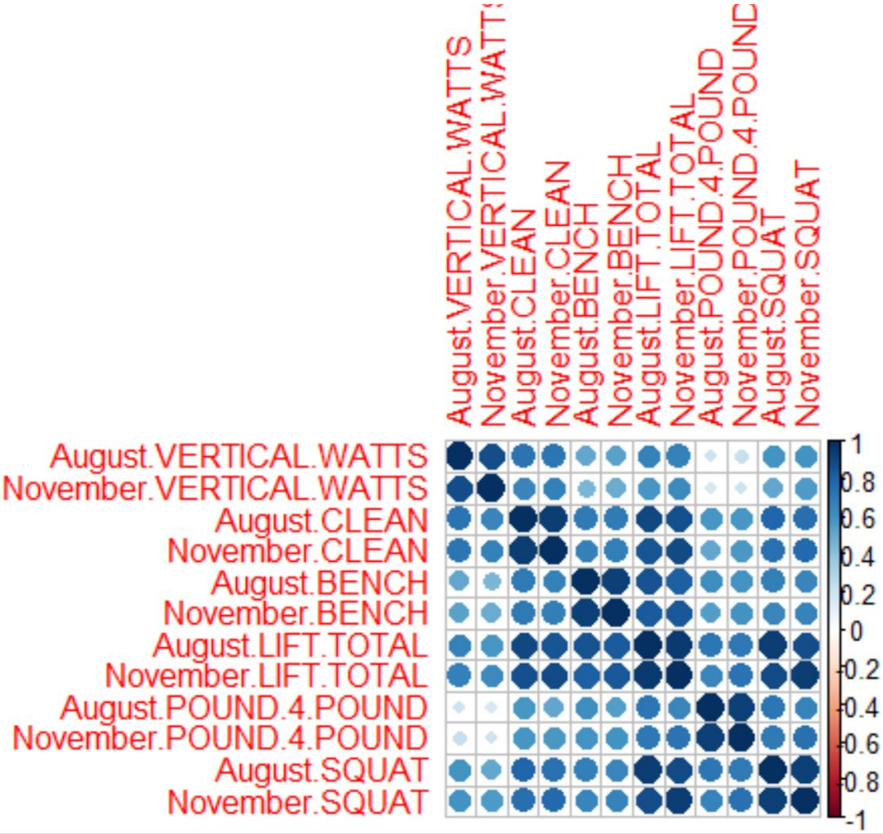
- **Assists** - Number of assists (any one direct pass by a player to a teammate who then scores a goal without having to dodge or evade an opponent)
- **Goals + Assists** - Sum of Goals Scored and Assists
- **Shots** - Number of shots taken on goal
- **Shot%** - Percentage of Scored Goals out of Shots
- **Goals while up** - Number of goals scored while the team was up a player (due to penalty)
- **Goals while down** - Number of goals scored while the team was down a player (due to penalty)
- **Ground Balls** - Number of ground balls retrieved (recorded when a ball changes possession during live-ball play)
- **Turnovers** - Number of turnovers (credited to a player when the player loses possession to the opponent)
- **Caused Turnovers** - Number of turnovers initiated (credited to a player when the player's positive, aggressive actions causes a turnover by the opponent)

Appendix II: The below is an example of the format in which the data was in when first received from Duke. Note the multiple rows and years for each player (Player 1 has two years worth of data) and the duplicated header rows (34 and 2), which made data processing somewhat tedious. Note that for this image, names were replaced in the original data.

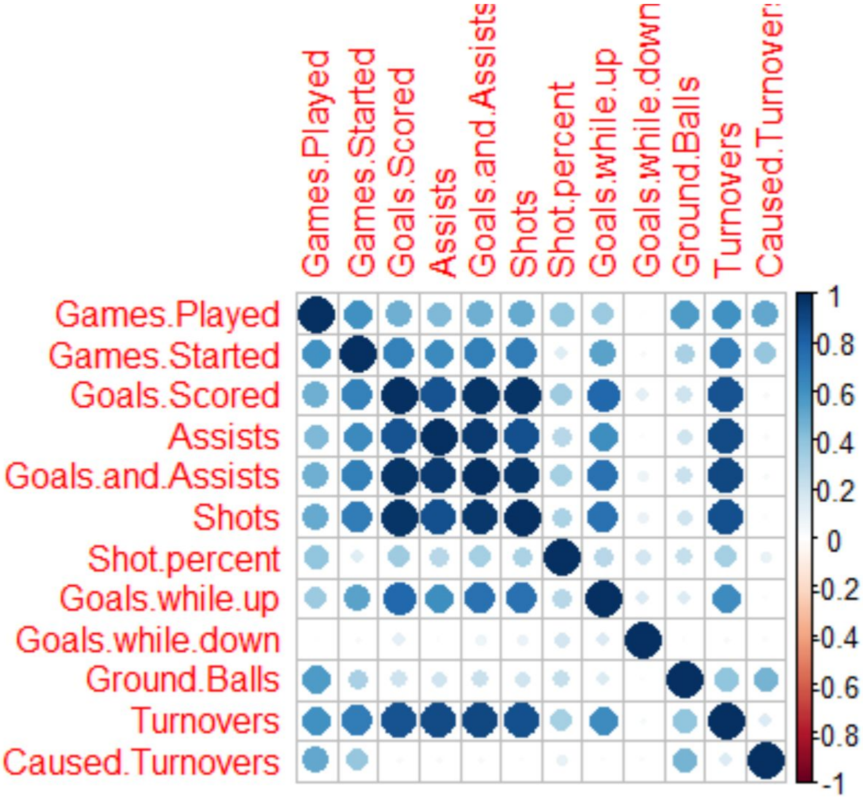


	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X
2	NAME	BODY WEIGHT	HEIGHT	WING SPAN	STANDING REACH	LONG JUMP	NO STEP VERTICAL	VERTICAL WATTS	PULLUPS REPS	VOLUME	CLEAN	BENCH	SQUAT	LIFT TOTAL	POUND 4 POUND	INVERTED T RIGHT	LEFT	5-10-5 SHUFFLE RIGHT	LEFT	5-10-5 RUN RIGHT	LEFT	SHOW UP %
3	[Player 1]																					
4	AUGUST 2015	174	6'0 3/4	6'2 1/2	7'9 1/2	9'0	28.5	5917.45	9	1814	FROSH	235	FROSH	235	1.35	5.76	5.91	4.88	4.97	4.45	4.35	
5	NOVEMBER 2015	177				9'0.5	29.0	6062.37	22	4508	230	240	360	830	4.67	5.63	5.64	5.04	4.94	4.52	4.46	100%
6	AUGUST 2016	187	6'0 1/2	6'3 1/4	7'9	9'1	27.0	5947.22	32	6858	265	255	370	890	4.76							
7	NOVEMBER 2016	185				9'1	29.5	6287.44	23	4878	270	255	380	905	4.9							89%
8																						
9	[Player 2]																					
10	AUGUST 2016	177	6'0	6'2 1/4	7'8 1/2	7'11	28.5	5968.84	8	1633	FROSH	205	FROSH	205	1.16							
11	NOVEMBER 2016	178				8'4	28.5	6005.93	11	2265	220	220	320	760	4.26							97%
12																						
13	[Player 3]																					
14	AUGUST 2013	197							9	2025	FROSH	245	FROSH	245	1.24							
15	NOVEMBER 2013	196	6'0 1/4						11	2478	210	250	285	745	3.80							100%
16	AUGUST 2014	196	6'0 1/2	6'4	7'11 1/4		24.5	5742.64	14	3123	230	260	305	795	4.06							
17	NOVEMBER 2014	193					26.5	5987.28	15	3300	235	255	320	810	4.21							98%
18	AUGUST 2015	193	6'0 1/2	6'4	7'11 3/4	8'6.5	26.5	5993.45	20	4406	240	250	355	845	4.38	6.04	5.94	4.99	5.03	4.42	4.54	
19	NOVEMBER 2015	196				9'5	29.0	6446.72	23	5143	275	285	365	925	4.72	INJ	INJ	INJ	INJ	INJ	INJ	100%
20	AUGUST 2016	208	6'0 1/2	6'4	7'11 3/4	INJ	INJ	INJ	19	4478	INJ	275	INJ	275	1.32							
21	NOVEMBER 2016	202				9'0	29.0	6565.93	20	4588	INJ	285	INJ	285	1.41							100%
22																						
23	[Player 4]																					
24	AUGUST 2015	176	5'10	6'0 1/2	7'8 1/2	8'0	29.0	6035.65	5	1018	FROSH	195	FROSH	195	1.11	6.10	6.19	4.96	5.06	4.63	4.54	
25	NOVEMBER 2015	174				7'10	32.0	6461.19	13	2623	200	190	250	640	3.67	5.89	5.86	5.16	5.32	4.50	4.57	92%
26	AUGUST 2016	181	5'10	6'0 1/2	7'7 1/2	8'0	29.5	6213.45	15	3128	225	225	315	765	4.23							
27	NOVEMBER 2016	185				8'5	29.5	6287.44	18	3818	230	235	350	815	4.42							97%
28																						
29	[Player 5]																					
30	AUGUST 2016	187	5'10 1/4	6'1 1/4	7'6	9'3	22.0	5172.21	6	1285	245	245	370	860	4.61							
31	NOVEMBER 2016	187				8'6	26.0	5801.26	21	4509	260	245	395	900	4.81							100%
32																						
33	BLUE DEVILS LACROSSE																					
34	NAME	BODY WEIGHT	HEIGHT	WING SPAN	STANDING REACH	LONG JUMP	NO STEP VERTICAL	VERTICAL WATTS	PULLUPS REPS	VOLUME	CLEAN	BENCH	SQUAT	LIFT TOTAL	POUND 4 POUND	INVERTED T RIGHT	LEFT	5-10-5 SHUFFLE RIGHT	LEFT	5-10-5 RUN RIGHT	LEFT	SHOW UP %
35	[Player 6]																					
36	AUGUST 2013	209							2	478	FROSH	255	FROSH	255	1.22							
37	NOVEMBER 2013	213	6'0 3/4						6	1459	235	260	330	825	3.87							100%
38	AUGUST 2014	211	6'0 3/4	6'1 1/4	7'10 1/4		25.5	6215.40	7	1670	240	285	345	870	4.12							
39	NOVEMBER 2014	204					24.5	5905.01	11	2541	250	295	370	915	4.50							100%
40	AUGUST 2015	200	6'1	6'1 1/4	7'10 1/4	8'4.5	27.5	6299.72	14	3188	245	300	405	950	4.75	5.96	5.94	4.69	4.63	4.62	4.54	

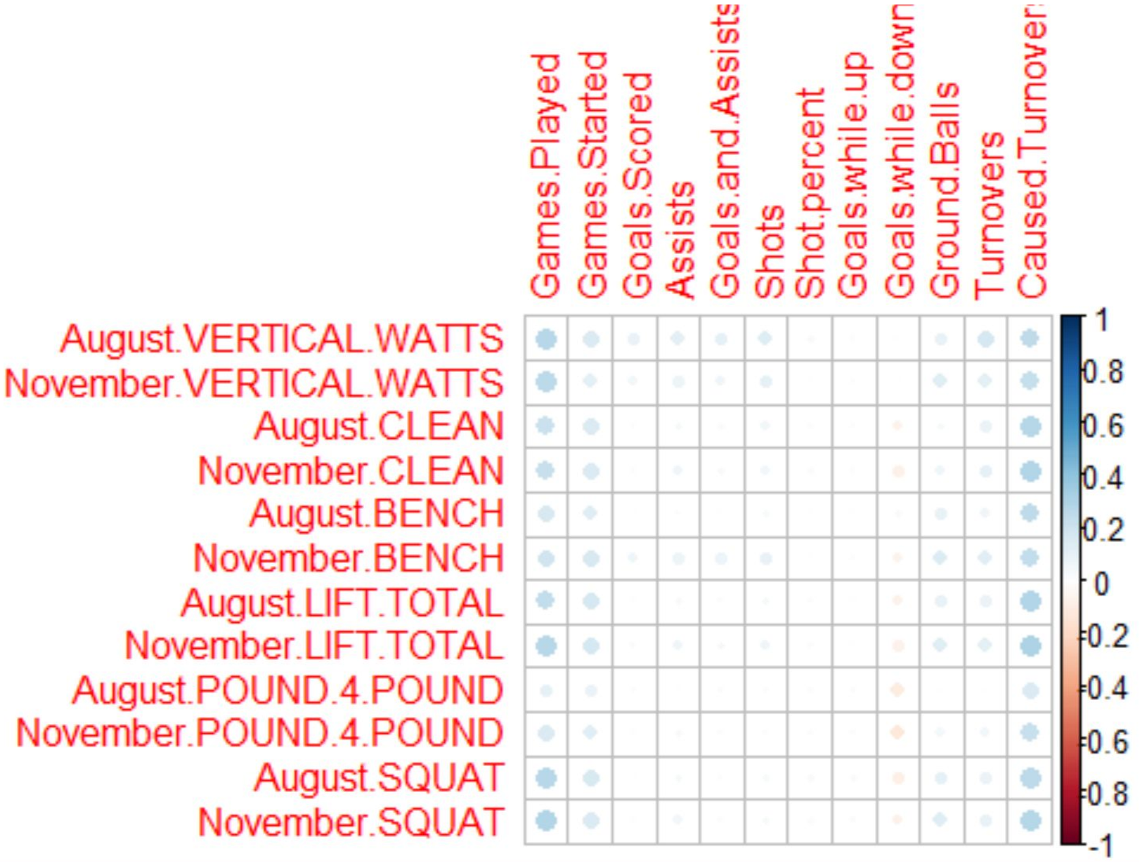
Appendix V: Correlation plot for weight and strength data



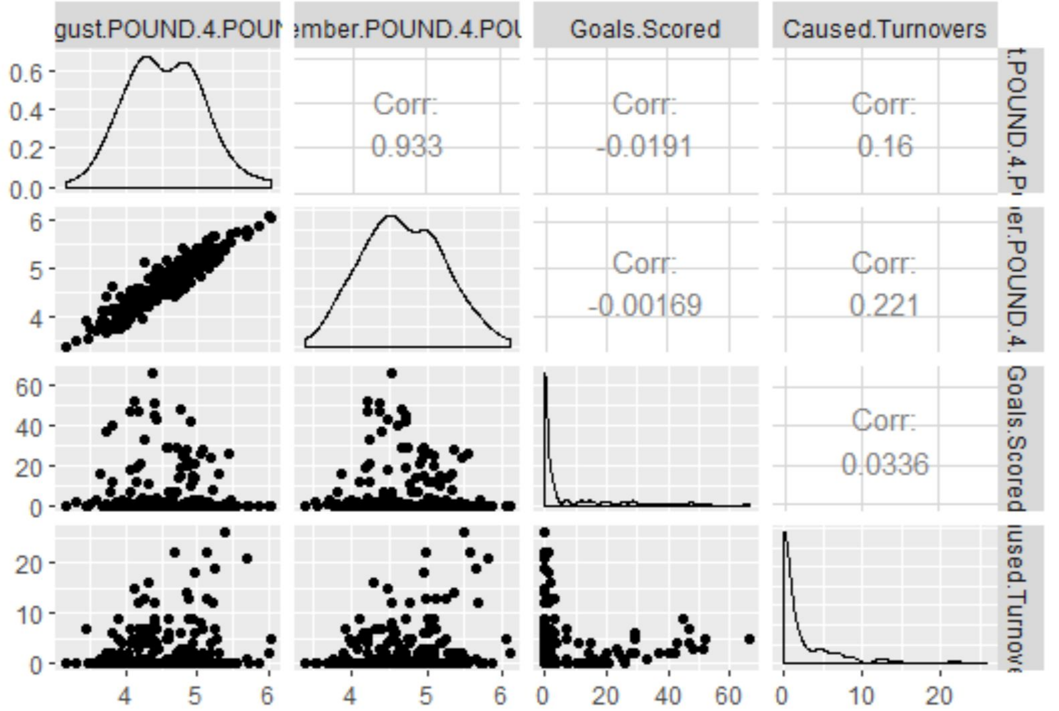
Appendix VI: Correlation plot for game performance data



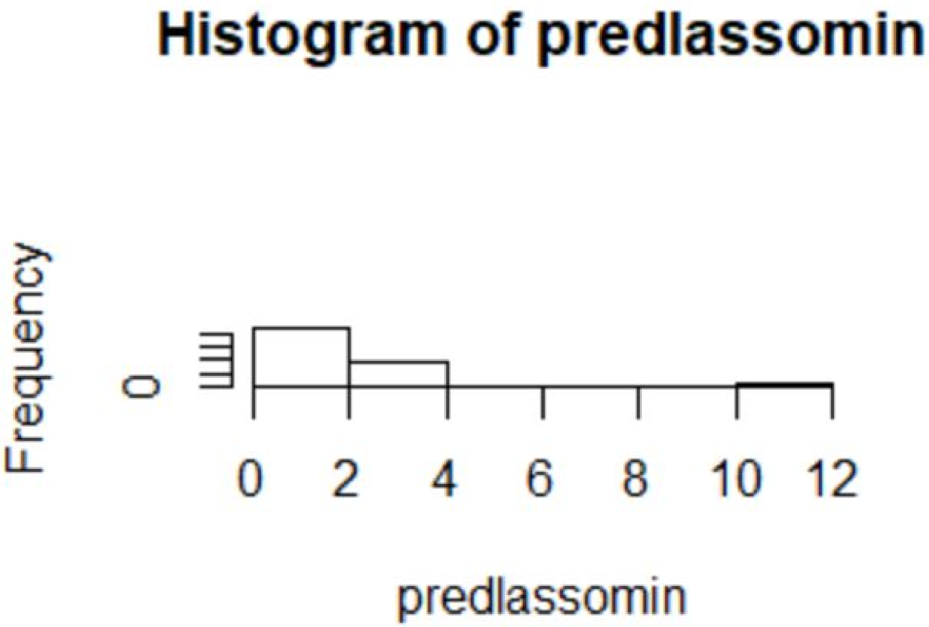
Appendix VII: Correlation plot for weight/strength data and game performance



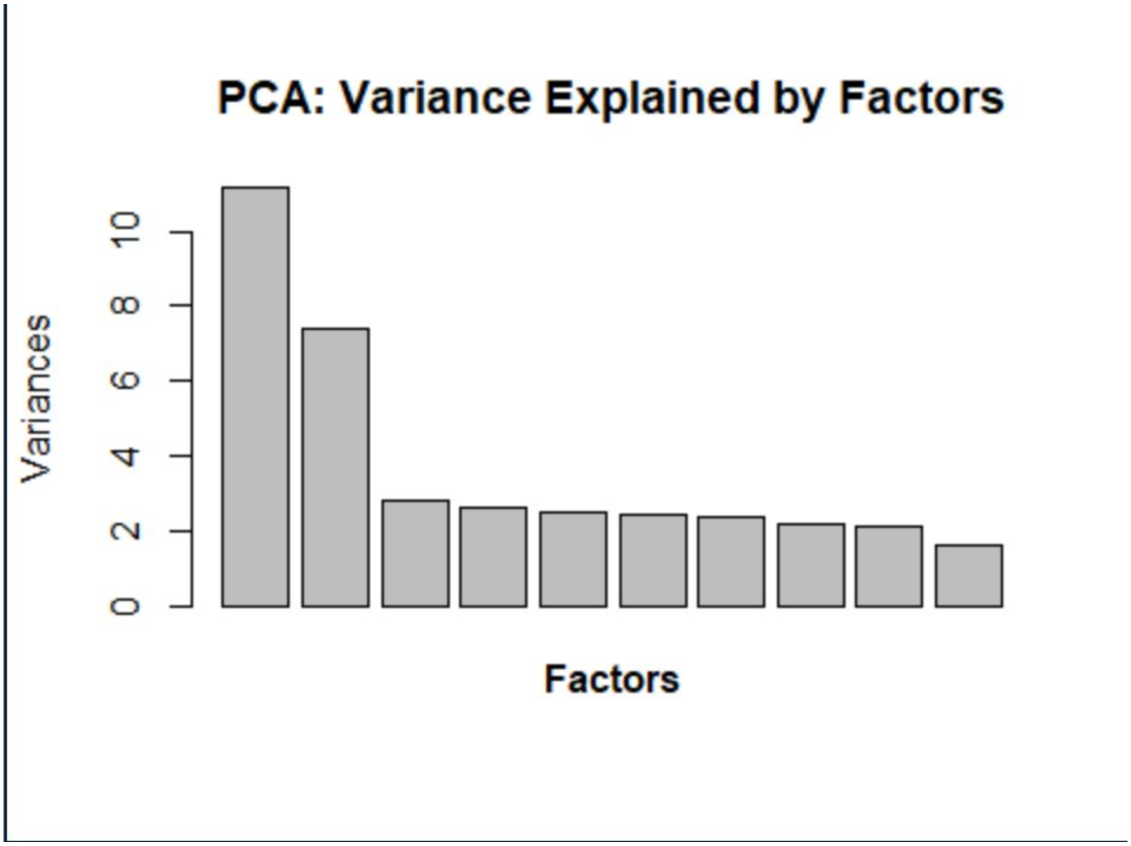
Appendix VIII: Correlation plots for Nov. and Aug Pound for Pound, Goals Scored and Caused Turnovers



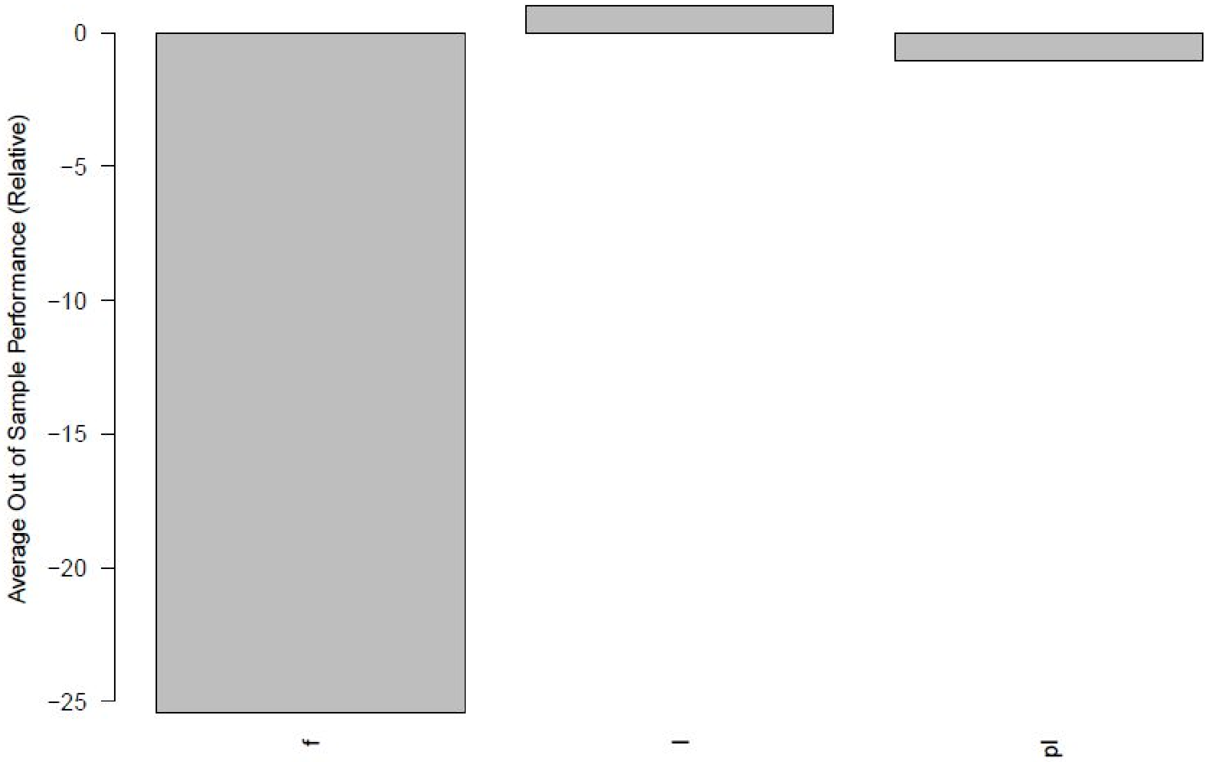
Appendix IX: Histogram of prediction of lasso minimum



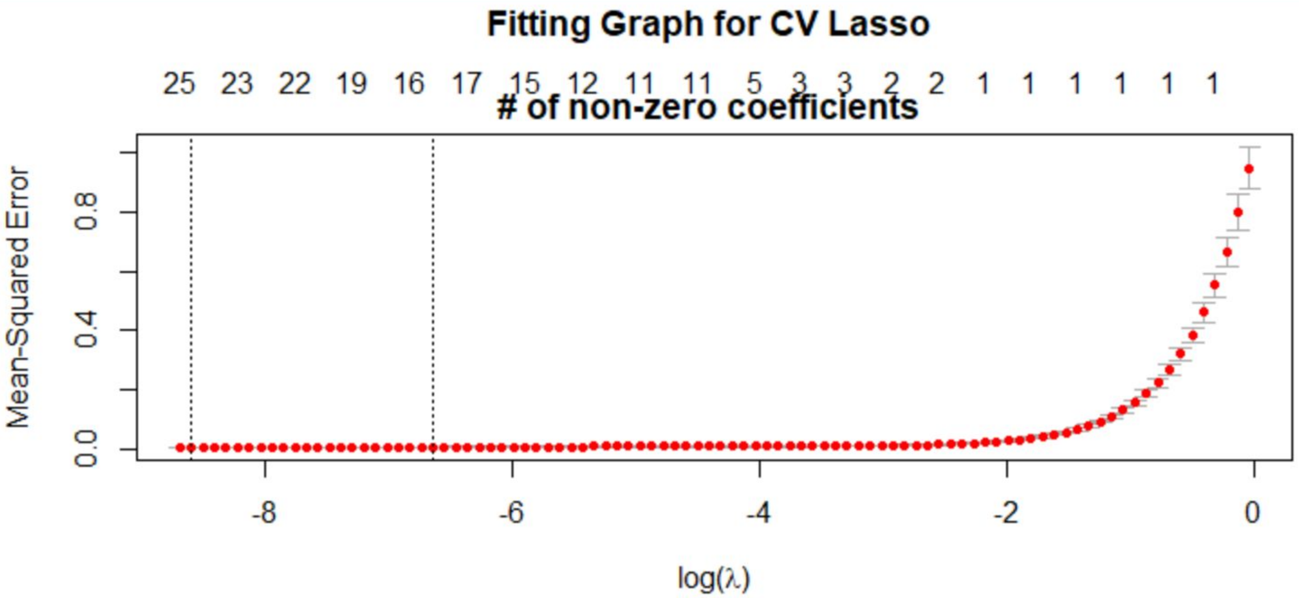
Appendix X: Results of PCA: shows that 2 sets of factors heavily explain much of the variance observed



Appendix XI: Model performance of the three models. From left to right the models are linear regression, lasso and post lasso.



Appendix XII: Fitting graph of cross validated lasso, showing the number of nonzero coefficients to be considered


Fitting Graph for CV Lasso
# of non-zero coefficients
Mean-Squared Error
log(λ)