CSC 642
Statistical Learning
May 7, 2024
Dr. Vanessa Aguiar-Pulido

What Tactics Should Coaches Prioritize in Soccer?

By Patrick Geraghty and Zidong Liu

1. Introduction

Soccer is the world's most popular sport, yet it has not felt the data revolution like in other sports such as baseball, basketball, or football. A career in soccer analytics has little stability. There is not a ton of sophisticated research in this area for various reasons. The game is hard to predict, upsets are common, and there is no true optimal play style. There is also of course resistance to change from traditional scouts and coaches. This in part makes this an attractive space, as there is a lot of room for growth on the research side. In this project we aim to build models to predict or determine the outcome of a game, with the goal of identifying the most important features, or metrics to get results in the beautiful game, specifically in Major League Soccer, the highest tier of soccer in the United States.

This is a classification problem, where we want to predict the outcome of a match, as mentioned previously. If successful, this could have big implications. There are many metrics of course that people find indicative of what team was superior in any game, but all of this is based on eye tests and qualitative feelings about the game. Possession is one of these metrics people often look at to identify the more dominant team in a game. With just a couple lines of code in R, you'll find that possession has extremely low predictive power alone. This project will allow us to back out which metrics were truly the most indicative in determining the final result. We will only be using performance related metrics in our models. This will be very valuable information as we will discover which of these metrics matter the most in determining the outcome of a game, which could then better inform tactical decision making. The model results also show us the true predictive power of many of these commonly used metrics in the micro field of soccer analytics. Furthermore, while there has been research done on the matter, as alluded to previously, none of this research has had a true imprint on the way the game is played. None of the research has been "groundbreaking" per se. Through our project and analysis, we will hope to have a fairly confident answer about what tactics coaches could be prioritizing.

2. State Of The Art

Soccer analytics is an emerging research field with several applications in a variety of fields. For example, there are models that attempt to predict' performance or the outcome of a game. This endeavor, aimed at forecasting match outcomes and identifying crucial metrics, is enriched by a broad number of scholarly contributions. The early exploration of passing sequence success rates by researchers in 1968[8] lays the groundwork for understanding the strategic elements that influence soccer outcomes. This is complemented by the nuanced analysis of passing patterns before and after goal scoring in the Premier League by Redwood-Brown [2]. He employed logistic regression models to discern the impact of passing patterns on the game's flow and outcome, highlighting the intricate relationship between tactical adjustments and match success. He emphasizes the tactical adjustments that can affect the game's flow and outcome.

Subsequently, the research extends to Apostolou and Tjortjis's insights[9] on the application of machine learning methods such as Random Forests and Support Vector Machines for predicting player positions and evaluating performance. This approach not only signifies a shift towards data-driven decision-making but also improves upon traditional methods by accommodating the nonlinear complexities of soccer data. Meanwhile, the examination of Big Data's role in tactical analysis in elite soccer[5] and the identification of machine learning-based methods for recognizing strong patterns[6] further showcase the evolving landscape of soccer analytics. These studies underscore the potential of advanced analytics in creating strategic models capable of accurately predicting game dynamics.

Furthermore, the application of Convolutional Neural Networks for analyzing spatial data from match footage represents a leap in tactical analysis[4]. This highlights the practical aspect of transforming raw data into actionable strategies. The comprehensive review of sports analytics algorithms for performance prediction[9], along with the detailed investigation into passing patterns before and after goal scoring in the FA Premier League[2] , collectively pave the way for a new era in soccer analytics. These seminal works, from the inception of analytical frameworks to the latest advancements in machine learning. By weaving these varied academic contributions into our analysis, our project not only aligns with the forefront of soccer analytics but also opens new avenues for using data to inform tactical decisions in Major League Soccer. This state-of-the-art compilation of literature not only demonstrates the historical depth and current trends in soccer data analysis but also sets a promising foundation for future explorations where strategy and analytics merge to redefine soccer's tactical paradigms.

3. Data

3.1 Dataset

The dataset we are using contains two rows, or observations, for every match during the 2022 and 2023 Major League Soccer regular season. The MLS is the highest tier of soccer in the United States. There are two rows per game as each instance has one of the two teams from the match as Team A and the other as Team B. Given this, we had 1,938 rows. Our data was also perfectly balanced. Each row contains a large number of event metrics (passing, shooting, possession, defensive actions, etc…), which are all quantitative, for both Team A and Team B, alongside a few variables concerning dates and venues. This gave us a total of 196 variables. There were plenty of redundant and other irrelevant variables we extracted and then performed feature engineering as well. There were no missing values. We got the data from the Football Reference website, which has tons of free sport event data to offer provided by Opta, or Stats Perform.

3.2 Data Collection

The data collection phase was very time consuming as it was a labor intensive process. We wanted a dataset with each game during the 2022 and 2023 MLS seasons containing detailed performance metrics for both teams. Football Reference has match logs for each team by season (FBRef, 2024). We were unaware of any way to download or scrape this data, thus we copy and pasted the match logs for every team into an excel sheet. There were fourteen different match logs we had to copy and paste for each team and season. There are different match logs for shooting, passing, pass types, shot and goal creation, defensive actions, possession, and miscellaneous, and we needed to get this same data for the opposing team as well.

3.3 Data Cleaning & Preparation

There was a lot of organization and data cleaning necessary to prepare our dataset for any models. We deleted many redundant variables from the different match logs. Then, we had to rename and label our variables. Each variable had a label to indicate whether it belonged to Team A or Team B. Beyond this, we created a few new features through different combinations of our original features. These can be found below in the appendix and are widely used metrics in soccer analytics. We deleted games that ended in a draw, which accounted for between 25-30% of our data set. We felt this would make our results more interpretable. We also deleted variables that were useless to our research purposes, such as the teams and others including goals. Finally, we scaled our predictors based on a minimum value of 9 and maximum value of 1 and split our data into a training and test set, leaving out 20% for testing. We used five k fold cross validation within our training set to ensure robust results. At the end of our data cleaning and preparation phase we were left with 1,410 observations and 203 variables. We did a significant amount of feature engineering, which is why the total number of variables did not lower..

We created a subset of predictors from our overall data set including metrics that coaches could prioritize. We did this for various reasons. First, we have many variables in our data set that are some linear combination of other variables. Second, there are plenty of metrics

that are relevant to the game, but cannot be used to inform tactical decision making. Had we done best subset selection on our entire data set, our final list of variables might not have been so interpretable. Nevertheless, we will create models using all the variables available as well as our subsetted list.

The application of convex optimization techniques further refines our approach to feature engineering and selection. By employing these methods, we harness powerful tools to solve optimization problems inherent in data preparation, especially when handling high-dimensional data sets. This approach not only enhances the efficiency of our feature selection process but also significantly reduces the computational complexity, allowing for a more streamlined analysis in identifying key predictors.

4. Methods

We used logistic regression and linear discriminant analysis to fit models with the sublist of predictors, and bagging and random forests to fit models with both sets. None of these models contained all variables as we excluded variables with information having to do with goals and expected goals, but for 'npxG/Shot.' The inclusion of those metrics would make it very difficult to extract and analyze best predictors. We elected to use bagging and random forests as best subset selection is performed within the learning process. A logistic regression, or LDA model would suffer from multicollinearity, random noise, and over-fitting without lessening the number of predictors. The only parameter we tuned in random forests was the number of features to be considered by each decision tree, which we set to be equal to half of the total number of predictors. These models allow us to assess the overall predictive power of the performance metrics, along with identifying the most influential individual variables.

Next, we tested our subset of metrics at a more secular level, by dividing the subset into two even smaller subsets of metrics based on whether or not the metric measures an in or out of possession principle. This could also be seen as attacking and defending, in a sense, but the terms in possession and out of possession are preferred. This is because an out of possession principle might be in order to create attacking danger. Ideally, we would have used goals for and against as the dependent variables for these two models in a regression approach. Due to our primary model predicting the result, we decided to stick with this to be able to compare all model results. And as mentioned, there are tactical principles relating to being without the ball that are for the purpose of creating attacking danger, so using either goals for or against as a dependent variable might not tell the full story. By doing this, we could see if any predictors changed in significance.

The last thing we checked was how much our prediction accuracy would go up if we included a few of the variables we omitted due to the question at hand. For example, we wished to see how much better the model would work if we included information on the home team. We also did this with expected difference, which is expected goals minus expected goals against in any given match. There was not much attention put on the results of individual predictors in these models, as the added variables take attention away from the metrics we wish to understand better. This exercise helped us better understand the limitations of our models, in terms of prediction accuracy.

Logistic Regression & Linear Discriminant Analysis
Logistic regression is a supervised machine learning classification method that models the probability of a certain event occurring, based on a set of input features. Linear discriminant analysis is a supervised machine learning classification method that makes predictions by using Bayes' limit theorem to determine the probability of the input variables belonging to either class.

Bagging & Random Forests
Bagging and random forests are two supervised machine learning methods that use decision trees as the basis for prediction. Bagging and random forests attempt to lower the high variance suffered by decision trees. These work by creating any number of bootstrapped training datasets in which the algorithm is trained on. A different decision tree is made for all

bootstrapped datasets and the prediction is the aggregation of all trees. Random forests' takes this one step further by only considering a random set of all the available predictors per tree. This is very helpful in cases where there are a few variables that are a lot more powerful than others, which is true in our case when including all possible variables.

Convex optimization
In addition to traditional statistical and machine learning models, the inclusion of convex optimization methods plays a crucial role in optimizing model parameters. These techniques are particularly effective in handling large-scale datasets and complex model structures, where they excel in finding the optimal solution by minimizing a convex loss function. This methodological enhancement is critical, especially when dealing with high-dimensional soccer analytics data, as it aids in refining our predictions and achieving higher accuracy.

To further refine our selection of input variables for convex optimization, we utilize the concept of convex hulls. A convex hull is the smallest convex set that encloses a given set of points in a multi-dimensional space, effectively determining the outer boundaries of possible outcomes for our input variables. By analyzing the convex hulls formed by various combinations of predictors, such as crosses and directness, we are able to identify the most influential variables that contribute to the robustness of our models. This geometric approach not only provides a visual representation of variable significance but also aids in selecting those metrics that encapsulate the most critical aspects of the game's dynamics. Utilizing these identified variables, we then apply convex optimization techniques to maximize the efficiency of our model, ensuring that the chosen predictors are optimally weighted to enhance predictive accuracy.

4.1 Evaluation

As this is a classification problem, we used our model's prediction accuracy to analyze, evaluate, and compare our model performance. We will primarily use accuracy given our classes are evenly distributed and we have no real incentive to avoid any specific occurance. To analyze and evaluate our best predictors, we will look at our coefficient weights for linear discriminant analysis. With logistic regression we will look for statistical significance. Lastly, we will use mean decrease in accuracy to evaluate the importance of the variables for bagging and random forests.

5. Results

Logistic Regression

Model	Training Accuracy	Testing Accuracy
Subset	76.95%	76.6%
In Possession*	72.96%	74.11%
Out of Possession*	68.44%	67.38%
Subset + Venue	80.13%	84.04%
Subset + xDif	81.3%	81.91%

Linear Discriminant Analysis

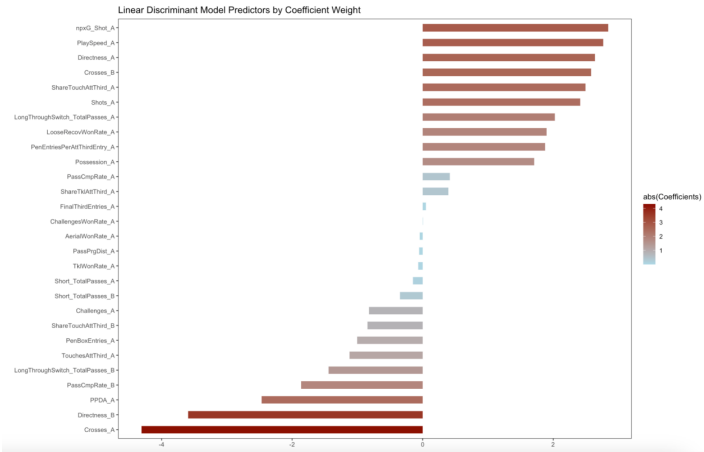
Model	Training Accuracy	Testing Accuracy
Subset	76.77%	78.37%
In Possession*	73.4%	75.17%
Out of Possession*	68.17%	67.38%
Subset + Venue	79.43%	85.1%
Subset + xDif	80.59%	82.98%

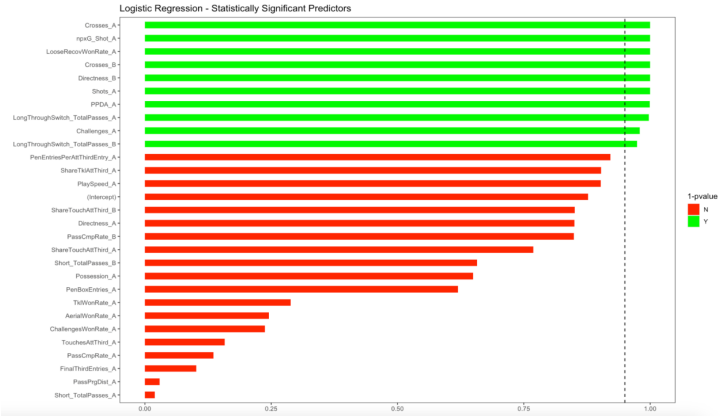
Bagging

Model	Training Accuracy	Testing Accuracy
All	79.88%	81.21%
Subset	74.02%	75.88%

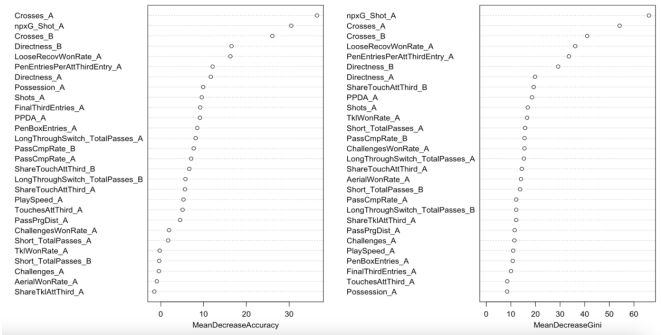
Random Forests

Model	Training Accuracy	Testing Accuracy
All	80.41%	81.56%
Subset	74.11%	73.76%

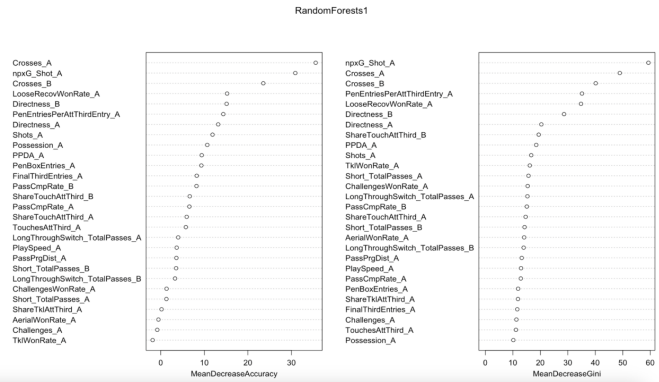




Bagging1



RandomForests1



6. Conclusions and Future Work

What tactics should coaches be prioritizing?
From the models we have produced, crosses (negative effect), opponent directness (negative effect), and expected goals per shot seem to be the strongest predictors. They are the only predictors that are statistically significant in logistic regression, and rank in the top five predictors based on weights of coefficients in our linear discriminant model and the mean decrease in accuracy in our bagging and random forests models. Crosses are usually seen as a positive action, so it may be of interest to some that its coefficient is negative. This could be for various reasons. It could be that teams cross the ball more when losing during a match, especially when there is not a lot of time remaining in games. On the other hand, it could be that many teams are playing compact low blocks in their defensive third, in which crossing the ball almost plays into the hand of the defense. That being said, there are moments when crosses are the right decision. Therefore, it might be smart to avoid "default' crosses, or crosses that are just lobbed balls into the box rather than directed at a teammate. Directness measures the overall share of the distance traveled from all of a team's completed passes that went forward toward the goal. The issue here is you could argue that there are two ways to minimize opponent directness. Based on the correlation matrix, which can be found in the appendix, it has quite a strong negative relationship with opponent pass completion rate. That means, in the last two seasons teams that perform well in directness do not usually perform at a high level in terms of their overall pass accuracy. Of course this is only a correlation, but it might mean the easiest way (not the best) to minimize opponent directness is through giving up the easy sideways or backwards pass. It is no surprise that expected goals per shot is here. The better average quality of your shots, the better chance you have of winning. This has already had an impact on the game to some degree. There has been a lot of casual research (blogs, social media posts, etc..) showing a decrease in the number of shots taken from outside of the box in a few popular leagues around the world.

Both opponent crosses (negative effect) and and loose ball recovery success rate are two of the top five predictors based on mean decrease in accuracy in both bagging and random forest models. Both of these were also statistically significant in our logistic regression model. Given the importance of crosses, it would be odd if opponent crosses were not relevant in our models. Many coaches emphasize the importance of winning your second balls. What this means is gaining possession of a loose ball after a header, a challenge, or any other type of action. Winning second balls comes down to a mixture of intelligent positioning, hustle, and determination. Lastly, passes per defensive action (negative effect) allowed is another predictor that was statistically significant in logistic regression, and had one of the higher weighted coefficients in our linear discriminant model. The lower the value, the more intensely a team presses without possession, making it more of a positive effect in actuality. Defensive pressure and intensity is a topic and tactic of the game that is seemingly becoming more and more popular every year. The idea is to defend aggressively in order to recover possession in more dangerous areas. This is very much a high risk, high reward type of deal, but it is definitely worth noting that many top teams in the game now try to defend with this type of intensity.

Our bagging and random forests models that included all variables, but for ones with goals and expected goals presented us some different findings. Shots on target, shots on target

rate, and surprisingly clearances for both teams were the most important predictors that were not included in our subset. One of the drawbacks to these models is the inability to know the sign of the effect of the predictor. We plotted two histograms together of clearances based on the result, and it actually seems as if the sign of the impact it has on our model is positive. It sounds somewhat far-fetched, but it could be that teams that are clearing the ball less are playing risky passes in their defensive third. Perhaps many teams were unsuccessful in doing so. At the same time, there is practically no correlation between clearances and expected goal difference, expected goals for, and expected goals against.

Adding information about which team was at home in every match increased our testing accuracy by a significant amount. It is definitely worth noting that our testing accuracy was much higher than our training accuracy. None of our other models experienced this same increase in accuracy during testing. Interestingly enough, the same did not happen when we included expected difference. Furthermore, the accuracy went up, but not as much as expected. The venue had more of an impact than expected goal difference in terms of improving our initial model's accuracy.

In our conclusion, the integration of convex optimization has significantly enhanced our understanding of optimal tactical choices in soccer analytics. By applying these methods, we have identified that the weight assigned to directness and passing accuracy (PPDA_A) should be optimized to enhance team performance. This analysis helps in pinpointing the precise tactical adjustments that could lead to a higher probability of winning matches.

Furthermore, the use of clustering techniques has allowed us to segment the teams into groups based on similar characteristics and performance metrics. This segmentation is based on a comprehensive analysis involving key performance indicators such as ball possession, shot quality, and defensive actions, which were processed through convex optimization to determine the most effective grouping strategy. This approach not only clarifies the comparative strengths and weaknesses of different teams but also aids in formulating tailored strategies against different clusters of opponents.

Our models suggest that focusing on high-quality shot creation and minimizing opponent directness are the most statistically significant predictors of match outcomes. These insights form the basis for recommending tactical shifts that prioritize these aspects of gameplay. As the next step in our research, we propose a deeper investigation into the application of convex optimization to refine these findings further and explore additional tactical elements that could influence game outcomes. Such advanced analytical techniques will pave the way for more nuanced and effective soccer strategies tailored to the unique dynamics of each match.

Model Analysis

In general, we achieved successful and consistent results. Our prediction accuracies are high given the limitations of predicting a soccer match. There is information we excluded to best answer our question that would of course be helpful, such as what team was home or away. Beyond this, our logistic regression and linear discriminant models were very similar in terms of accuracy and the importance of specific predictors. The exact same can be said for our bagging and random forests models. There was plenty of overlap, adding to the strength of the results and conclusions possible from this analysis. It is no surprise that our models

with nearly all of the variables outperformed the subset, but the subset is not much worse, and is much more interpretable. The logistic regression and linear discriminant models that were separated by either in or out of possession produced very similar results to the bigger model in terms of statistical significance and the weights of coefficients. This is why there is no real mention of these models in the section above.

Future Work and Investigation

Some further research should be conducted on some of the metrics that were most relevant across our different models. This way a more confident conclusion can be made about the impact these predictors have on the outcome of a match. Specifically, seeing on average when it is that most teams who lose a game are crossing the ball most. if there is no clear indication of an increase in crossing when down, then a more precise conclusion can be made about crosses.

The importance of clearance in our bigger models was definitely a surprise. There are definitely some grounds to look further into why this might have been the case. Perhaps some sort of linear combination between clearances and defensive actions could be tested.

Another area to further investigate are loose ball recoveries, especially where they happen. It might be that loose ball recovery success rate in a specific area of the field could be very vital. it could also be that lots of high quality chances are created from loose ball recoveries in a certain zone.

Convex optimization also extends its utility to the model evaluation phase, where it contributes to optimizing the validation process. By integrating convex optimization strategies, we can rigorously assess the stability and robustness of our predictive models. This not only ensures that our models perform consistently across different datasets but also improves their generalization capabilities, a vital aspect when forecasting soccer match outcomes.

Shortcomings and Drawbacks

It is more difficult to predict or extract valuable insight from data pertaining to Major League Soccer. This boils down to the salary cap. It is widely known that player wages are the best predictor of team success. In the MLS, all teams are subject to a salary cap, but are allowed three players that they can pay however much they like. Thus, the variation in player wages in the MLS is only a fraction of what it is in other leagues globally. This lower variation in wages likely leads to lower variation in performance related metrics, thus making it harder to predict match outcomes in the MLS.

Not all tactical principles are directly measurable. With the level of data available to us, we can really only look at broader topics such as defensive pressure, ball control, or the nature of possession. Beyond this, few tactical principles can be perfectly tracked or measured, and in the case it is possible, more specific event data would be required.

Every opponent is different. The results from these models may not necessarily apply to a specific team, and therefore must be taken with a grain of salt. In the real world, you would want to model, or assess, every single opponent separately.

The scoreline at any point in time likely impacts tactical strategy to at least some extent. That being said, teams might perform a specific action more so when losing, or more so when winning, which could then in turn make that action appear as one that helps in getting good results, when this might not be the case.

The approach we took in making the subset of variables was a hard one, because we were essentially doing our own best subset selection based on what we thought was most appropriate, which of course adds bias. But, given the data, and the question at hand, we got to a point where it felt necessary to do such a thing. Given our question and problem, creating a model that was easy to analyze and interpret was of the utmost importance.

7. References

[1]Fbref. (2024.). *Football Statistics and History*. FBref.com. https://fbref.com/en/

[2]Redwood-Brown, A. (2008). Passing patterns before and after goal scoring in FA Premier League Soccer. *International Journal of Performance Analysis in Sport*, *8*(3), 172–182. https://doi.org/10.1080/24748668.2008.11868458

[3]Collet, C. (2012). The possession game? A comparative analysis of ball retention and team success in European and international football, 2007–2010. *Journal of Sports Sciences*, *31*(2), 123–136. https://doi.org/10.1080/02640414.2012.727455

[4]Matthew G. S. Kerr. (2015). Applying Machine Learning to Event Data in Soccer. Master's thesis, MASSACHUSETTS INSTITUTE OF TECHNOLOGY, June 2015. http://hdl.handle.net/1721.1/100607

[5]Rein, R., & Memmert, D. (2016). Big data and tactical analysis in elite soccer: future challenges and opportunities for sports science. *SpringerPlus*, *5*(1). https://doi.org/10.1186/s40064-016-3108-2

[6]Geurkink, Y., Boone, J., Verstockt, S., & Bourgois, J. G. (2021). Machine Learning-Based Identification of the Strongest Predictive Variables of Winning and Losing in Belgian Professional Soccer. *Applied Sciences*, *11*(5), 2378. https://doi.org/10.3390/app11052378

[7]Pino-Ortega, J., Rojas-Valverde, D., Gómez-Carmona, C. D., & Rico-González, M. (2021). Training Design, Performance Analysis, and Talent Identification—A Systematic Review about the Most Relevant Variables through the Principal Component Analysis in Soccer, Basketball, and Rugby. *International Journal of Environmental Research and Public Health*, *18*(5), 2642. https://doi.org/10.3390/ijerph18052642

[8]C. Reep and B. Benjamin. (1968). Skill and chance in association football. Journal of the Royal Statistical Society. Series A (General), 131(4):581–585, January 1968. https://www.researchgate.net/publication/271760194_Skill_and_Chance_in_Ball_Games

[9]K. Apostolou and C. Tjortjis, "Sports Analytics algorithms for performance prediction," 2019 10th International Conference on Information, Intelligence, Systems and Applications (IISA), Patras, Greece, 2019, pp. 1-4, doi: 10.1109/IISA.2019.8900754. https://doi.org/10.1109/IISA.2019.8900754

8. Appendix

Final Variable List

Variable	**Description**	**Calculation (if necessary)**
Result_A	win, draw, or loss	
xDif_A	the expected score difference	expected goals - expected goals conceded
Venue_A	if team A was home or away in the given match	
Gls_A	goals scored	
Ast_A	goals assisted	
Gls/Shot_A	goals per shot	goals/shots
Gls/SOT_A	goals per shot on target	goals/shots on target
Shots_A	shots taken	
ShotsOnTarget_A	shots taken on target	
ShotOnTargetRate_A	share of shots on target	shots on target/shots
AvgShotDist_A	average distance to goal per shot	
ShotsFK_A	shots from free kicks	
xG_A	expected goals (every shot has an xG value from 0-1 based on the probability of scoring from said shot)	sum of the xGs for every shot
npxG_A	non penalty expected goals	xG-xG from penalties
npxG_Shot_A	npxG per shot	npxG/shot
G-xG_A	goals minus expected goals	goals-xG
PassCmp_A	pass completions	
PassAtt_A	pass attempts	
PassCmpRate_A	pass completion rate	pass completions/pass attempts
PassTotDist_A	total distance of all passes	
PassPrgDist_A	total progressive distance of all passes	
ShortPassCmp_A	short pass completions	

ShortPassAtt_A	short pass attempts	
ShortPassCmpRate_A	short pass completion rate	short pass completions/ short pass attempts
MedPassCmp_A	medium pass completions	
MedPassAtt_A	medium pass attempts	
MedPassCmpRate_A	medium pass completion rate	medium pass completions/medium pass attempts
LongPassCmp_A	long pass completions	
LongPassAtt_A	long pass attempts	
LongPassCmpRate_A	long pass completion rate	long pass completions/long pass attempts
xAG_A	expected goals assisted	
xA_A	expected assists (probability of any pass being an assist, averaging all passes)	
KP_A	passes leading to a shot	
SuccPassToFinThird_A	completed passes to the final, or attacking third	
SuccPassIntoBox_A	completed passes into the box	
SuccCross_A	successful crosses	
SuccPrgPass_A	progressive passes	
ThroughBall_A	passes into space behind the defensive line	
Switches_A	passes that travel over 40 yards the width of the field	
Crosses_A	cross attempts	
Corners_A	corner kicks	
SCA_A	shot creating actions (the two actions preceding any shot)	
PassLiveSCA_A	live pass SCA	
PassDeadSCA_A	pass from dead ball SCA	
TakeOnSCA_A	SCA from take ons	

ShotSCA_A	SCA from shots	
FoulSCA_A	SCA from fouls	
DefActionSCA_A	SCA from defensive actions	
Tkl_A	tackles	
TklW_A	tackles where possession was won back	
TklDefThird_A	tackles in the defensive third	
TklMidThird_A	tackles in the middle third	
TklAttThird_A	tackles in the final third	
ChallengesWon_A	challenges where possession was regained	
Challenges_A	opponent dribblers challenged	
ChallengesWonRate_A	share of challenges where possession was regained	challenges won/challenges
Int_A	interceptions	
Clr_A	clearances	
ErrorShot_A	errors leading to opponent shot	
Possession_A	the % of the game spent with the ball	
Touches_A	every instance of an attacking player in possession of the ball	
TouchesDefPen_A	touches in their own box	
TouchesDefThird_A	touches in the defensive third	
TouchesMidThird_A	touches in the middle third	
TouchesAttThird_A	touches in the final third	
TouchesAttPen_A	touches in the penalty box	
TakeOn_A	take ons attempted	
SuccTakeOn_A	successful take ons	
SuccTakeOnRate_A	take on success rate	take ons/successful take ons

Carries_A	instances of player running with the ball	
CarryTotDist_A	total carrying distance	
CarryPrgDist_A	total progressive carrying distance	
PrgC_A	progressive carries	
CarryFinThird_A	carries into the final third	
CarryPenBox_A	carries into the box	
Miscontrol_A	miscontrols of the ball	
Dispossessed_A	losing possession of the ball	
CrdY_A	yellow cards	
CrdR_A	red cards	
Fls_A	fouls committed	
LooseBallRecov_A	loose ball recoveries	
AerialWon_A	aerial duels won	
AerialWonRate_A	aerial duel success rate	aerial duels won/aerial duels
PPDA_A*	passes per defensive action allowed, measuring pressing intensity (a lower value means more defensive intensity)	opponent pass attempts/tackles+interceptions+clearances+fouls committed
FieldTilt_A*	a measure of spatial field dominance	touches in the final third/touches in the final third+opponent touches in the final third
Directness_A*	a measure of how direct a team is in possession of the ball	progressive passing distance/total passing distance
LongThroughSwitch_TotalPasses_A*	a measure of how often a team plays long	long pass attempts+through balls+switches/pass attempts
Short_TotalPasses_A*	a measure of how often a team plays short	short pass attempts/pass attempts
ShareTklDefThird_A*	share of tackles in defensive third	tackles in defensive third/tackles
ShareTklMidThird_A*	share of tackles in middle	tackles in middle

	third	third/tackles
ShareTklAttThird_A*	share of tackles in attacking third	tackles in attacking third/tackles
TklWonRate_A*	share of tackles where possession was regained	tackles won/tackles
LooseRecovWonRate_A*	share of loose balls recovered	loose ball recoveries/loose balls
ShotsAssistedRate_A*	share of shots that were assisted	key passes/shots
FinalThirdEntries_A*	entries into attacking third	passes into final third + carries into final third
PenBoxEntries_A*	entries into penalty box	passes into penalty box + carries into penalty box
PenEntryPerAttThirdEntry_A*	entries into penalty box per entries into attacking third	penalty box entries/final third entries
PlaySpeed_A*	a measure of how quickly a team passes when in possession	pass completions/share of possession

***the variables with an asterisk were manually engineered*

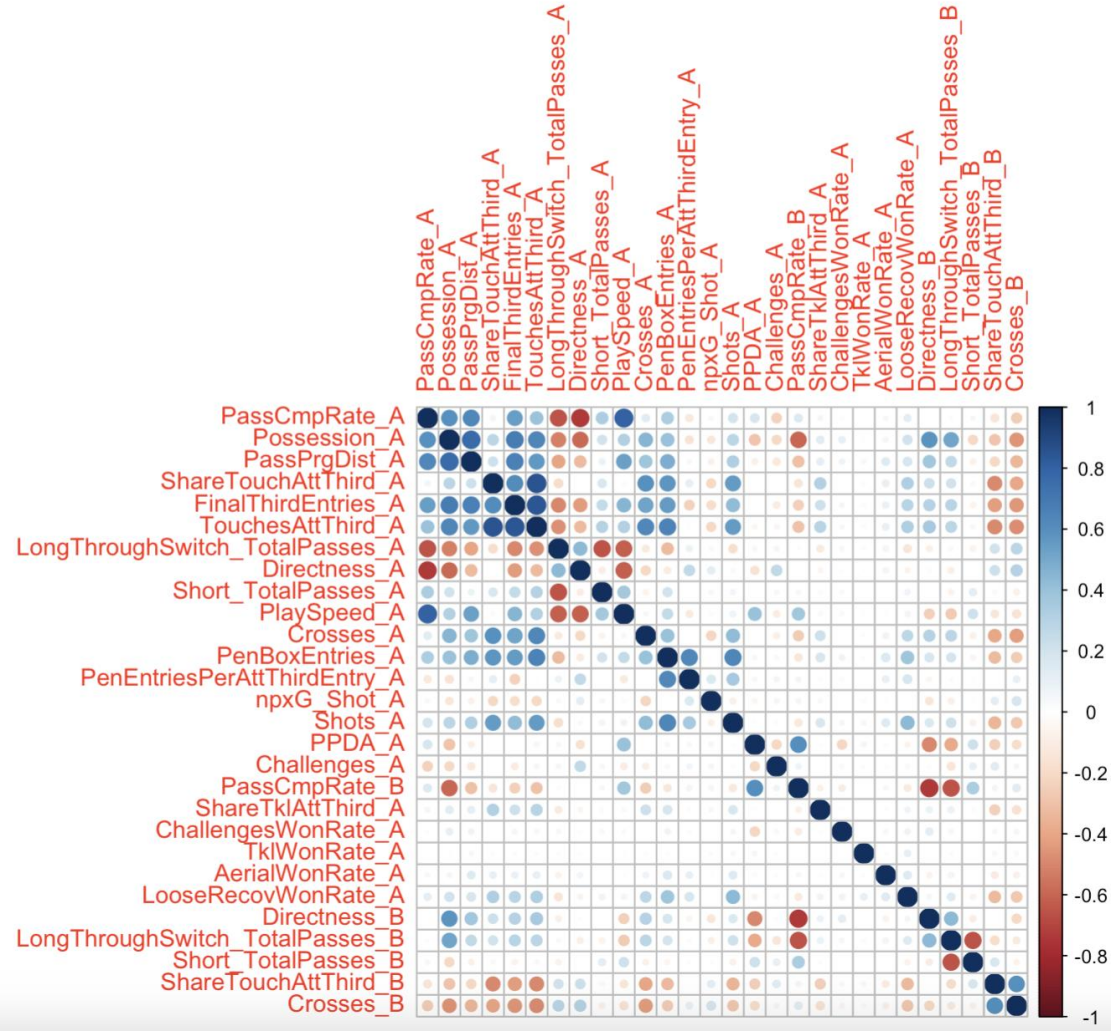
the same variables are also present for Team B

Subset of Variables Measuring Tactical Principles Coaches Could Prioritize

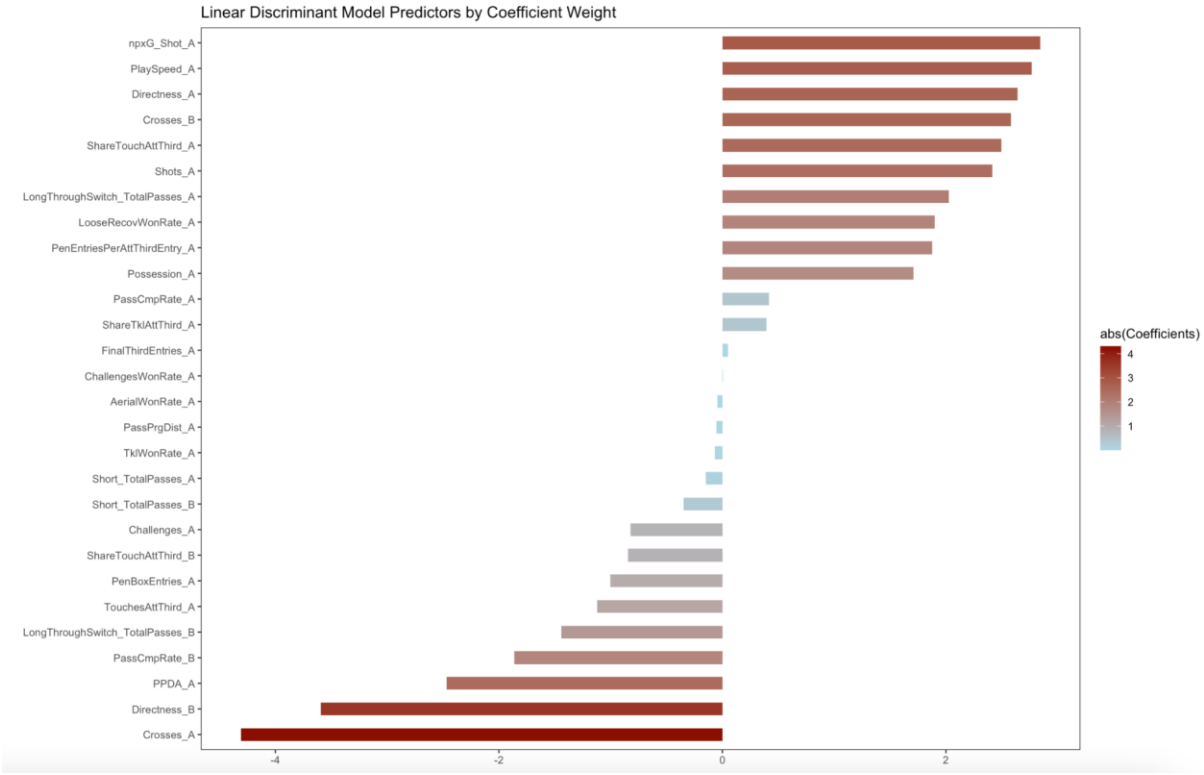
Variable	**In/Out of Possession**	**Tactical Principle**
PassCmpRate_A	In	control
Possession_A	In	control
PassPrgDist_A	In	progression
ShareTouchAttThird_A	In	progression
FinalThirdEntries_A	In	progression
TouchesAttThird_A	In	progression
LongSwitchThrough_TotalPasses_A	In	nature of build up
Directness_A	In	nature of build up
Short_TotalPasses_A	In	nature of build up
PlaySpeed_A	In	nature of build up
Crosses_A	In	play in att third

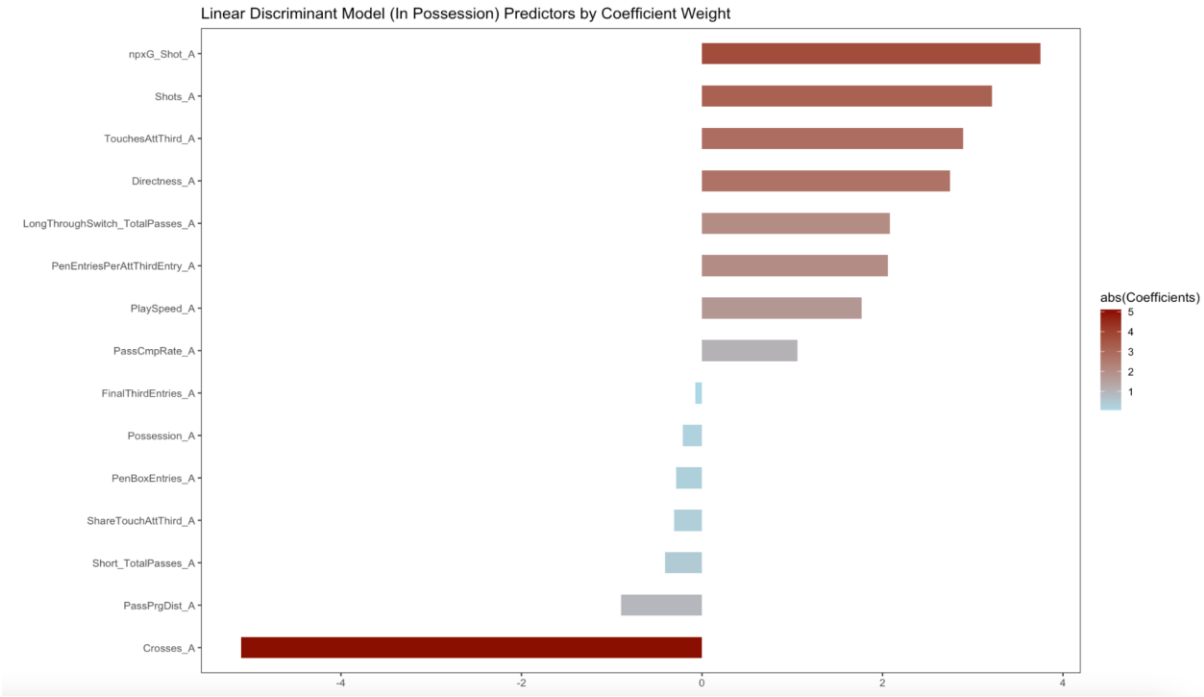
PenBoxEntries_A	In	play in att third
PenEntriesPerAttThirdEntry_A	In	play in att third
npxG_Shot_A	In	play in att third
Shots_A	In	play in att third
PPDA_A	Out	defensive intensity
Challenges_A	Out	defensive intensity
PassCmpRate_B	Out	defensive intensity
ShareTklAttThird_A	Out	high press
ShareTouchAttThird_B	Out	high press
ChallengesWonRate_A	Out	duels
TklWonRate_A	Out	duels
AerialWonRate_A	Out	duels
LooseRecovWonRate_A	Out	duels
Directness_B	Out	forcing opponent nature of build up
LongThroughSwitch_TotalPasses_B	Out	forcing opponent nature of build up
Short_TotalPasses_B	Out	forcing opponent nature of build up
Crosses_B	Out	opponent play in att third

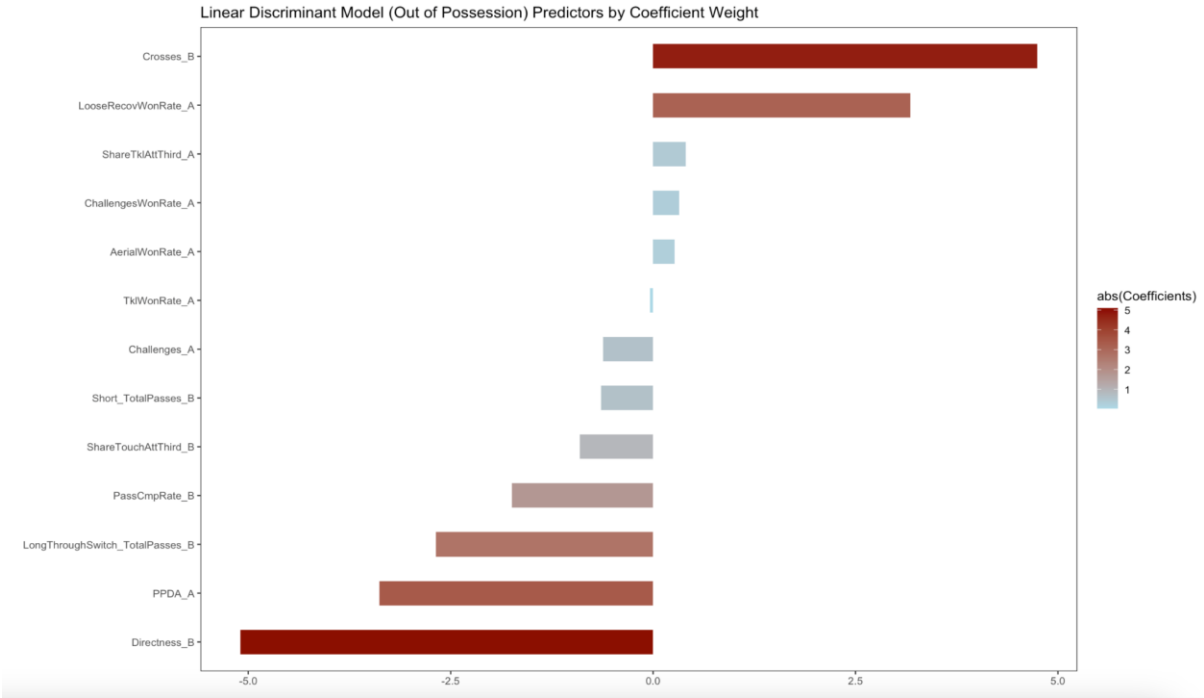
Correlation Matrix for Variables Above

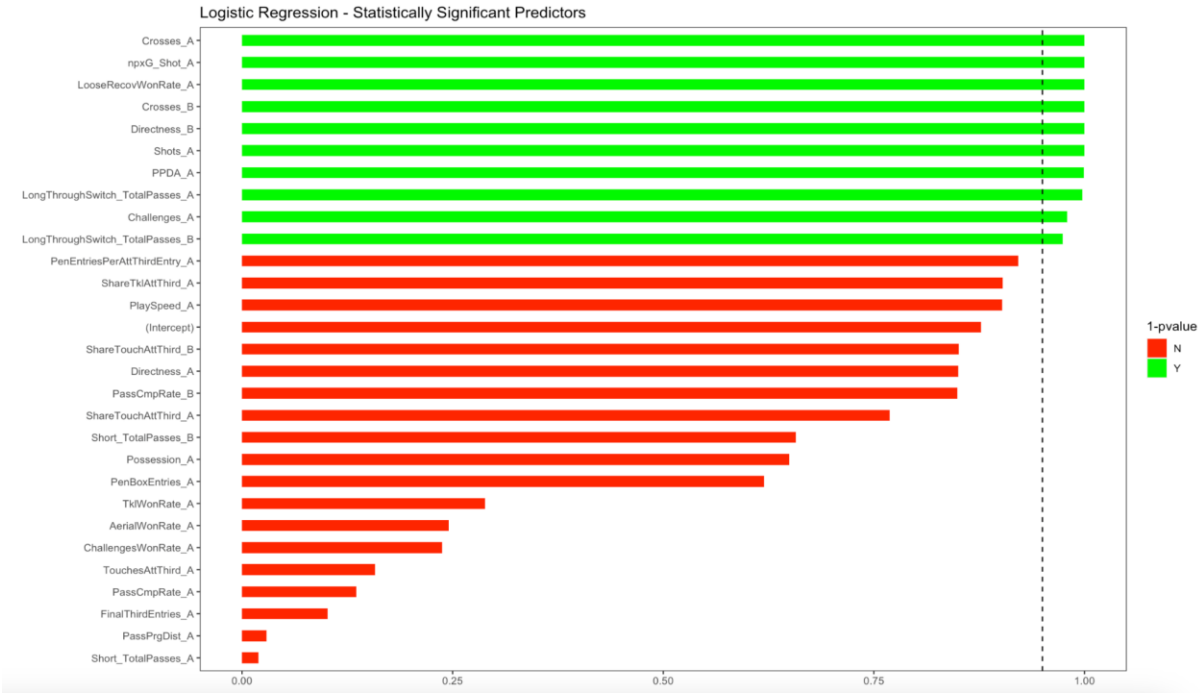


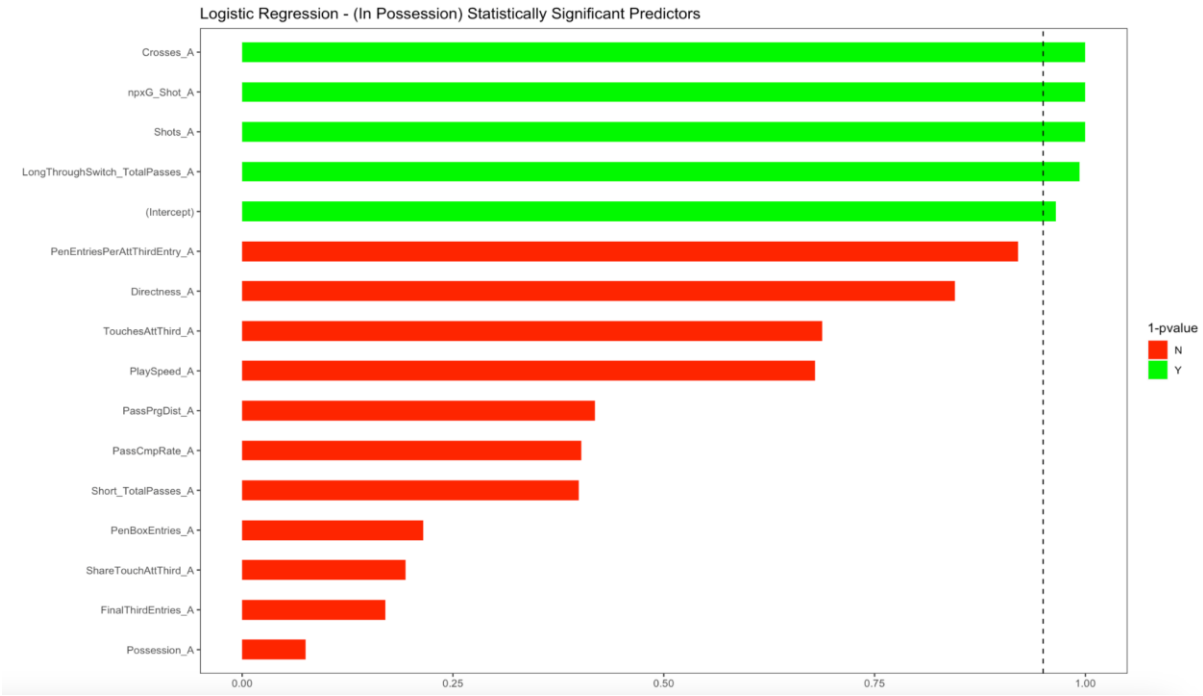
Plots of Model Results

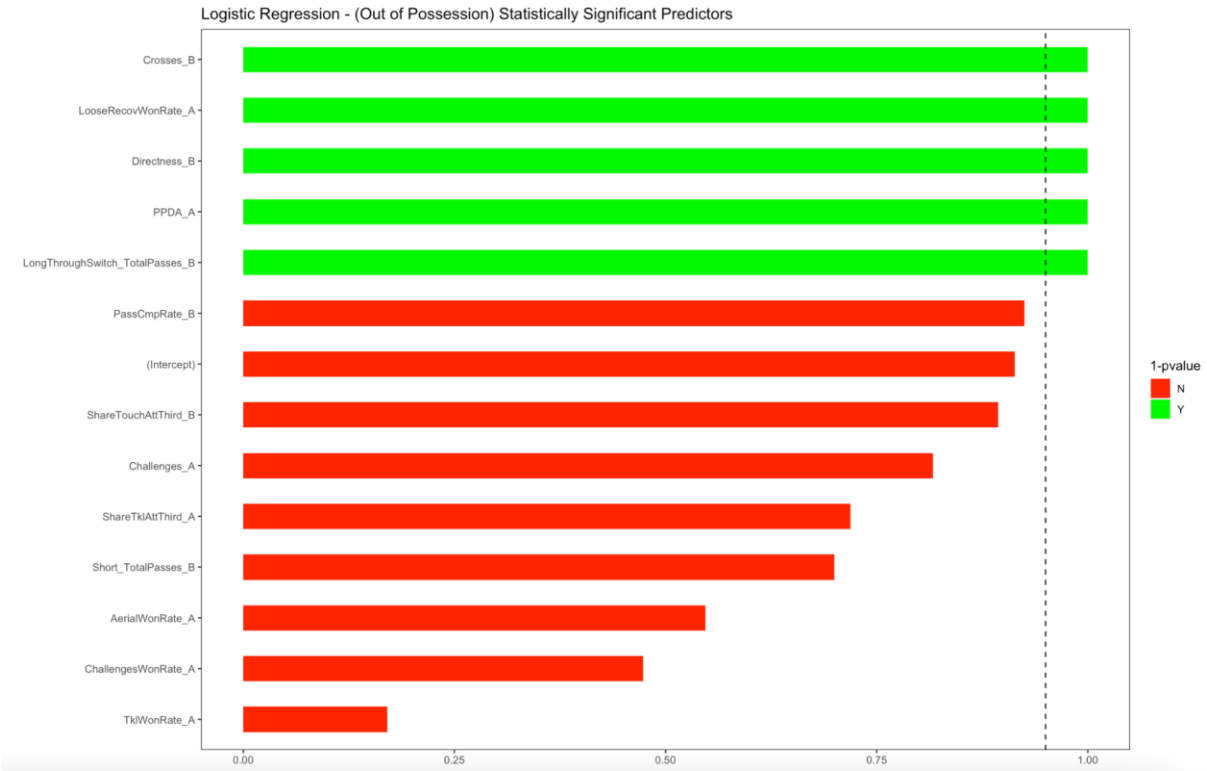

Linear Discriminant Model Predictors by Coefficient Weight
npxG_Shot_A
PlaySpeed_A
Directness_A
Crosses_B
ShareTouchAttThird_A
Shots_A
LongThroughSwitch_TotalPasses_A
LooseRecovWonRate_A
PenEntriesPerAttThirdEntry_A
Possession_A
PassCmpRate_A
ShareTklAttThird_A
FinalThirdEntries_A
ChallengesWonRate_A
AerialWonRate_A
PassPrgDist_A
TklWonRate_A
Short_TotalPasses_A
Short_TotalPasses_B
Challenges_A
ShareTouchAttThird_B
PenBoxEntries_A
TouchesAttThird_A
LongThroughSwitch_TotalPasses_B
PassCmpRate_B
PPDA_A
Directness_B
Crosses_A
-4
-2
0
2
abs(Coefficients)
4
3
2
1


Linear Discriminant Model (In Possession) Predictors by Coefficient Weight
npxG_Shot_A
Shots_A
TouchesAttThird_A
Directness_A
LongThroughSwitch_TotalPasses_A
PenEntriesPerAttThirdEntry_A
PlaySpeed_A
PassCmpRate_A
FinalThirdEntries_A
Possession_A
PenBoxEntries_A
ShareTouchAttThird_A
Short_TotalPasses_A
PassPrgDist_A
Crosses_A
-4
-2
0
2
4
abs(Coefficients)
5
4
3
2
1

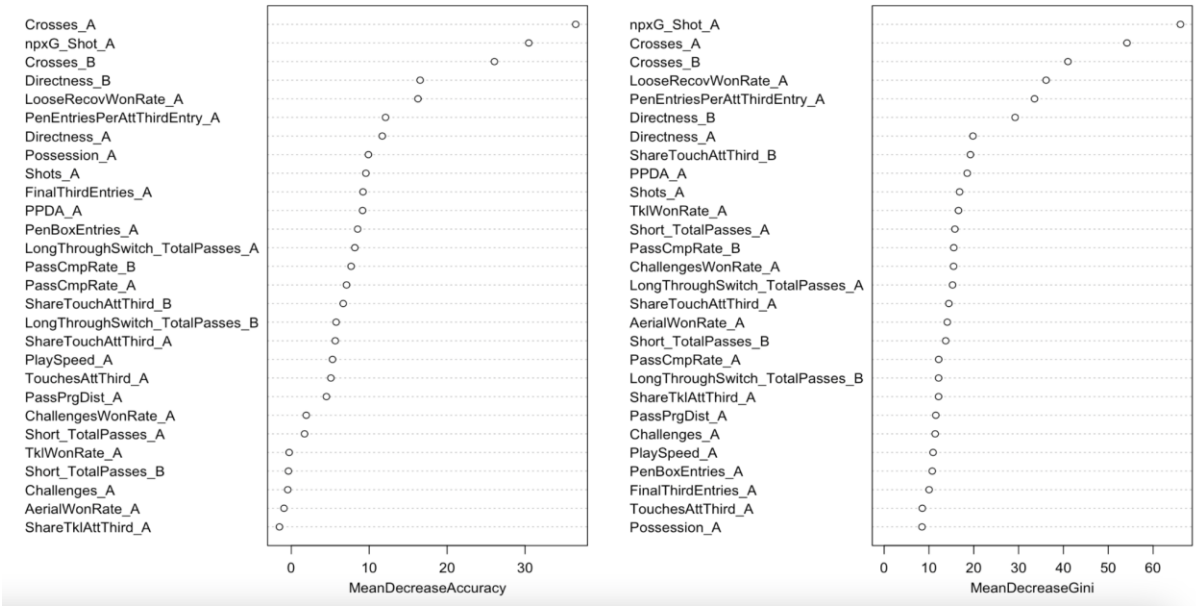

Linear Discriminant Model (Out of Possession) Predictors by Coefficient Weight
Crosses_B
LooseRecovWonRate_A
ShareTklAttThird_A
ChallengesWonRate_A
AerialWonRate_A
TklWonRate_A
Challenges_A
Short_TotalPasses_B
ShareTouchAttThird_B
PassCmpRate_B
LongThroughSwitch_TotalPasses_B
PPDA_A
Directness_B
-5.0
-2.5
0.0
2.5
5.0
abs(Coefficients)
5
4
3
2
1


Logistic Regression - Statistically Significant Predictors
Crosses_A
npxG_Shot_A
LooseRecovWonRate_A
Crosses_B
Directness_B
Shots_A
PPDA_A
LongThroughSwitch_TotalPasses_A
Challenges_A
LongThroughSwitch_TotalPasses_B
PenEntriesPerAttThirdEntry_A
ShareTklAttThird_A
PlaySpeed_A
(Intercept)
ShareTouchAttThird_B
Directness_A
PassCmpRate_B
ShareTouchAttThird_A
Short_TotalPasses_B
Possession_A
PenBoxEntries_A
TklWonRate_A
AerialWonRate_A
ChallengesWonRate_A
TouchesAttThird_A
PassCmpRate_A
FinalThirdEntries_A
PassPrgDist_A
Short_TotalPasses_A
0.00
0.25
0.50
0.75
1.00
1-pvalue
N
Y

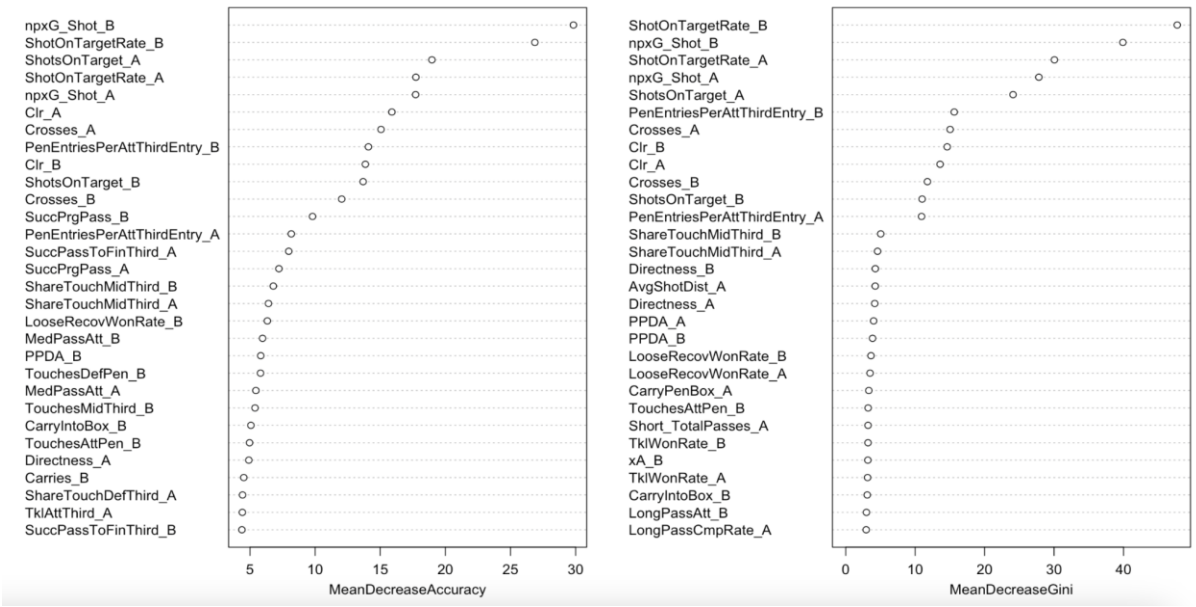

Logistic Regression - (In Possession) Statistically Significant Predictors
Crosses_A
npxG_Shot_A
Shots_A
LongThroughSwitch_TotalPasses_A
(Intercept)
PenEntriesPerAttThirdEntry_A
Directness_A
TouchesAttThird_A
PlaySpeed_A
PassPrgDist_A
PassCmpRate_A
Short_TotalPasses_A
PenBoxEntries_A
ShareTouchAttThird_A
FinalThirdEntries_A
Possession_A
0.00
0.25
0.50
0.75
1.00
1-pvalue
N
Y


Logistic Regression - (Out of Possession) Statistically Significant Predictors
Crosses_B
LooseRecovWonRate_A
Directness_B
PPDA_A
LongThroughSwitch_TotalPasses_B
PassCmpRate_B
(Intercept)
ShareTouchAttThird_B
Challenges_A
ShareTklAttThird_A
Short_TotalPasses_B
AerialWonRate_A
ChallengesWonRate_A
TklWonRate_A
0.00
0.25
0.50
0.75
1.00
1-pvalue
N
Y

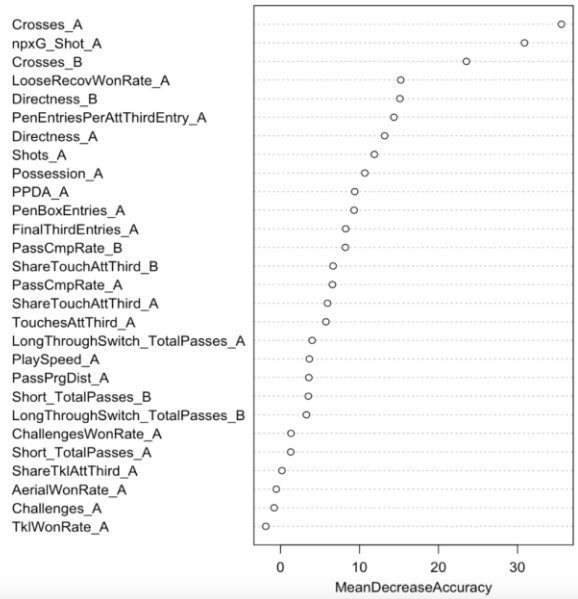
Bagging1

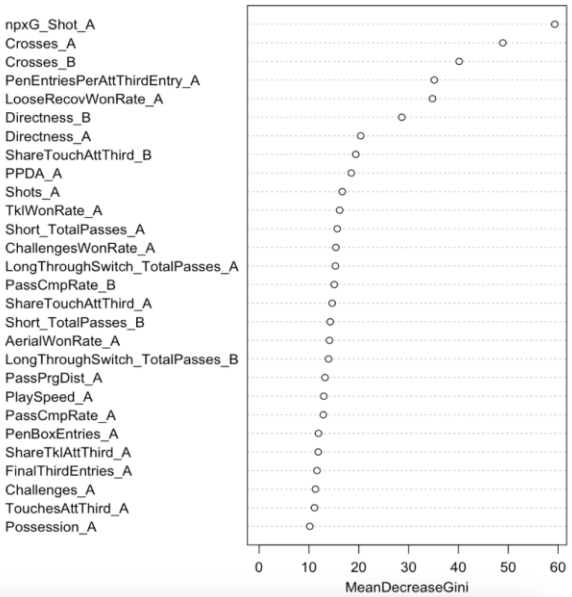


Bagging2

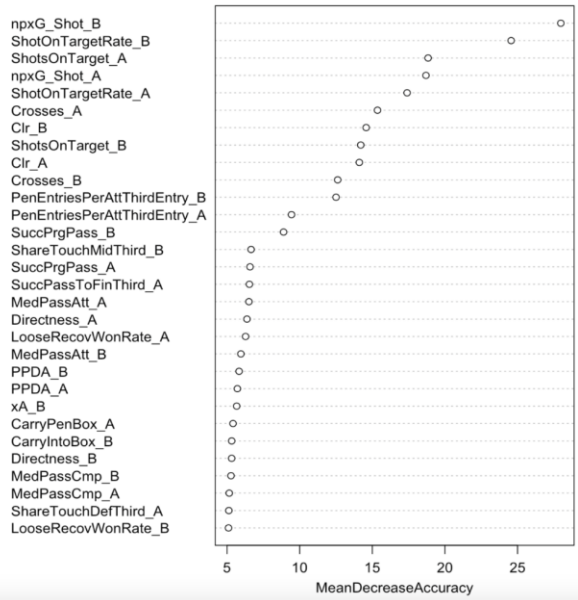


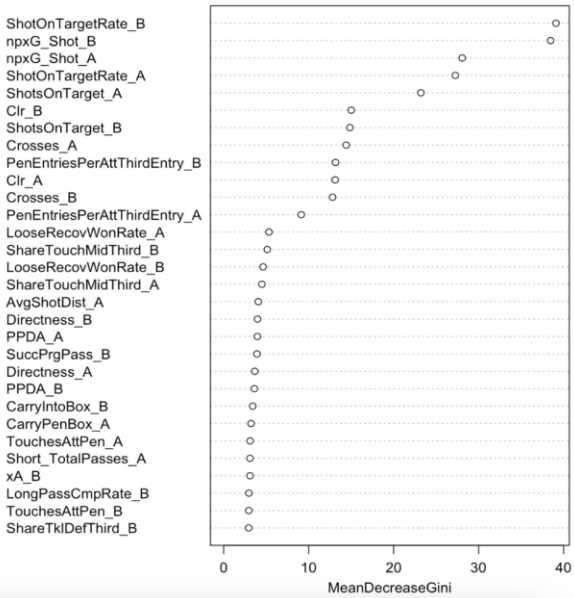
RandomForests1

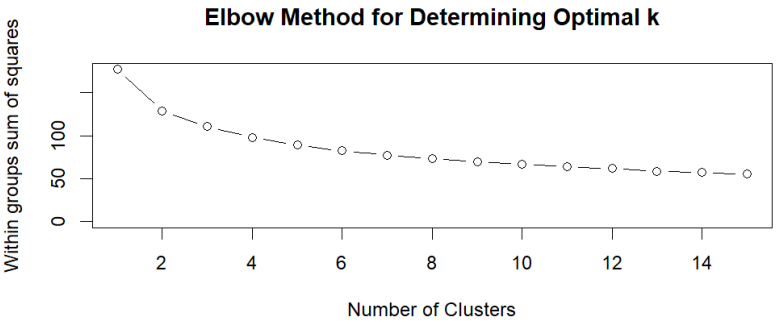


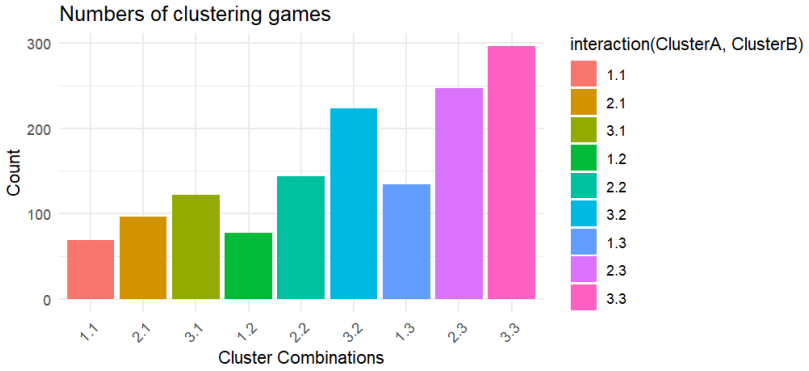


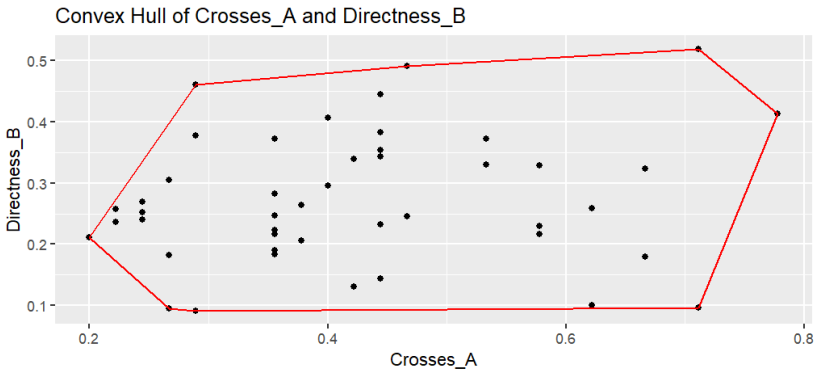
RandomForests2

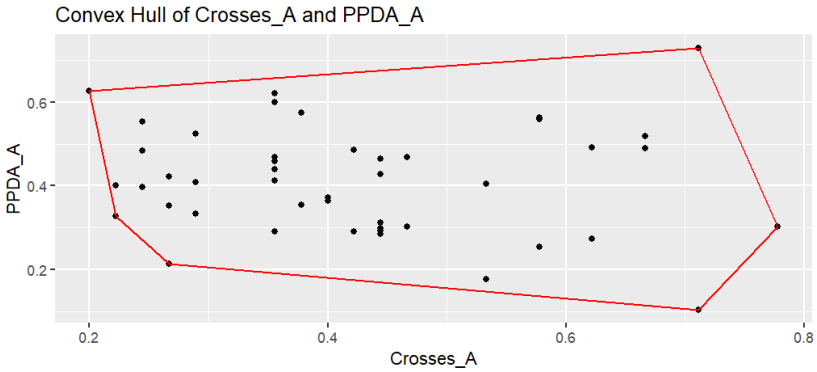













```
> print(optimized_values)
         Crosses_A        Directness_B             PPDA_A FinalThirdEntries_A         npxG_Shot_A
        0.35555028          0.21636368          0.29138820          0.09755588          0.65517280
>
> optimal_value <- result$value
>
> print(-optimal_value)
[1] 1.398574
```