Discerning Important Statistics in ACC Soccer

Instruction

The Athletic Coast Conference is an athletic conference consisting of fifteen teams based on the eastern side of the United States. Founded in 1953, its members compete in a variety of different sports such as football, basketball, lacrosse, field hockey, and golf. Twelve of the fifteen teams in the ACC sponsor soccer. Split into two divisions of six, these teams compete amongst each other each year for seeding in the ACC Men's Soccer Tournament. Based on the results of the regular season and the conference tournament, teams then qualify for the NCAA Tournament, where they compete against teams from other conferences to claim the ultimate prize in men's college soccer, the NCAA National Championship.

Lit Review

Although my research question will focus specifically on predicting matches inside the realm of ACC soccer, plenty of work has been done on the sport at large. The application of machine learning to soccer is relatively nascent in nature. The nature of the sport does not lend itself to mathematical principles; there are too many variables in soccer that are changing every minute as the game develops and the teams react to the dynamic flow. Ben Ulmer and Matthew Fernandez, two computer scientists at Stanford University, attempted to predict the results of soccer matches in the English Premier League using five different classifiers including but not limited to: naive Bayes, hidden Markov model, support vector machines, and random forests. The random forest algorithm was among the best methods in their study, resulting in an training

error rate of about .045 and a test error rate of about .5. They found that every model used underpredicted draws, showing the difficulty in differentiating ties from wins and losses respectively (Ulmer et al. 2013). Two university professors in Germany used World Cup data from 2002-2014 to test the predictive performance of random forests in comparison to more traditional regression methods. They concluded that, in general, random forest methods outperform the regression-based approaches (Schauberger et. al 2018). Ordered random forests were used to predict results in the Bundesliga, the first division of German soccer, to relatively good results (Goller et al. 2018).

Data

As a student at the University of Notre Dame, I work with the Notre Dame Sports Performance Team, specifically focusing on statistics projects in men's soccer. Through contacts on the soccer team and the Sports Performance Team, I gained access to an InStat database detailing ACC soccer games over the last 3 years. InStat is an online sports performance analysis service that works with FC Barcelona, Real Madrid, Chelsea, and many more of the world's top soccer clubs. They allow teams to conduct deep statistical analysis utilizing their massive database and unique InStat Scout program. The database I acquired includes all matches for all ACC teams in the 2016 and 2017 seasons. Each observation is a record of one player's performance in a specific game. It includes some important qualitative information about each game, such as the date it was played and the teams involved. The rest of the fields are detailed statistics regarding the performance of the players in the game, including data on passing,

shooting, and crossing. It also includes defensive statistics such as interceptions, ball recoveries, and more. An example of a few rows of data is shown below.

SURNAME	TEAM	MATCH	MATCH DATE	PLAYERS POSITION	INSTAT INDEX	GOALS	ASSISTS
Habrowski Matt	Notre Dame Fighting Irish	Notre Dame Fighting Irish 4:0 UC Irvine Anteaters	8/27/2016	CD	144	0	0
Habrowski Matt	Notre Dame Fighting Irish	Notre Dame Fighting Irish 1:0 New Mexico Lobos	8/28/2016	CD	142	0	0
Habrowski Matt	Notre Dame Fighting Irish	Team 1 0:5 Notre Dame Fighting Irish	9/4/2016	CD	147	0	0
Habrowski Matt	Notre Dame Fighting Irish	Notre Dame Fighting Irish 4:1 Virginia Tech Hokies	9/10/2016	CD	193	0	0
Habrowski Matt	Notre Dame Fighting Irish	Notre Dame Fighting Irish 1:0 Connecticut Huskies	9/14/2016	CD	166	0	0
Habrowski Matt	Notre Dame Fighting Irish	Louisville Cardinals 1:0 Notre Dame Fighting Irish	9/17/2016	CD	143	0	0
Habrowski Matt	Notre Dame Fighting Irish	Notre Dame Fighting Irish 2:1 Syracuse Orange	9/24/2016	CD	115	0	0

I have access to another dataset including goalkeepers, but, due to fundamental differences in that position relative to all other positions on the field as well as my own research interests, I will not use that dataset in this analysis.

Design

In this analysis, I will use the random forest algorithm in order to discern the most important statistics relating to winning, tying, and losing games in the ACC. There are many different positions in soccer, and each position possesses different requirements and responsibilities to contribute to positive results. If I attempted to run an analysis on every position, I could write a forty page paper. However, I want to focus on certain positions that I am personally interested in and that I believe hold the most impact on results in the sport, subject to natural limitations of the data. I have decided to analyze one role from every level of the structure of a team: defenders, defensive midfielders, attacking midfielders, and attackers. Central defenders are the most important component of every backline; they are responsible for protecting the central area in front of the goalie while starting attacks from deep inside their own half. I will analyze defensive midfielders whose main duties include protecting the space in front of the defense, winning the ball back from the other team, and distributing the ball to playmakers higher up the field. I also wanted to make sure I analyzed the wide areas; thus, for attacking

midfield, I will focus on the wingers that stretch the field and create dangerous attacking situations for their team. Lastly, I will analyze the strikers who just put the ball in the back of the net. For each position I have outlined, I will run the random forest algorithm in an attempt to isolate the most important statistics for each position that contribute to results in ACC soccer, and I will compare those results to a logistic regression comparing wins and losses and the effects of these statistics on these classes.

Data Preparation

For the best possible analysis of ACC soccer, I will discard the games that include any out-of-conference teams. This takes the total number of records down from 6951 to 3079. However, I believe this will ensure consistency and integrity of results by ignoring games that may have been played against significantly weaker opposition in order to improve team fitness or overall resume. The original dataset did not include the result of each game; it merely included a match column that described the result of each game. Using some Excel commands and an R script, I split the match column into the home team, the away team, the home team score, and the away team score. I then matched the column that denoted the team of the player in question with either the home team or away team column. Lastly, I used the score columns to identify the result of the game for each team, and thus, for each player. Also, I realized that minutes were an important factor in accumulating statistics for players; players with more minutes would be more likely to have higher counts merely due to the fact that they played more time. In order to remedy this, I used the minutes played column to put each raw statistic, ignoring percentages, in per-90 minute terms to ensure consistent comparison among players. Before actually running the

random forest algorithm, some of the qualitative columns need to be removed, such as player ID, kit number, name, etc. In addition, after some exploratory analysis, I realized that it was extremely difficult to accurately model ties in the data; there just aren't enough differentiating characteristics between ties and the other two classes. It turns out that ties only comprise about 20% of the data, so I eliminated those and focused my analysis on classifying wins and losses in ACC play.

Method

In this paper, I will be running the random forest algorithm to predict results based on different statistics for different positions. The random forest algorithm has a few defining characteristics. First of all, it is an ensemble method, which means it combines multiple types of models in order to improve and create the best possible predictions for your problem. It is a supervised learning method, which means that it utilizes data that is already classified or tagged and attempts to use that tag to learn something about the data.

The basic structure of random forests come from the decision tree method. In this procedure, the algorithm partitions the data iteratively by placing each observation into one of two groups where each group describes a particular predicted response. It continues this process of splitting the data until it cannot create a better split, or some pre-specified condition is reached, such as depth of tree or Gini index threshold. A few variations of the decision tree method exist, such as conditional inference trees which split based on p-values at each level of the tree. Decision trees hold many benefits in the world of machine learning; they are very easy to interpret and aesthetically pleasing, they are not very affected by outliers that plague

real-world data, and they automatically include interaction effects which are quite prevalent in large datasets. However, they also possess some established disadvantages. Decision trees are quite unstable; a change in prediction at one level of the tree can propagate down the rest of the tree, creating massive ripple effects centered around even seemingly insignificant changes. Collinearity among predictors also presents an issue when splitting the data. Lastly, decision trees result in a relatively poor predictive performance compared to other methods in machine learning due to, among other factors, the very rigid structure of the prediction function that is applied to the data to partition each observation.

The random forest method was designed to mitigate some of these inherent disadvantages in decision trees. At its core, the algorithm takes a bootstrap sample of a certain size from the data and creates a decision tree on that sample using a pre-specified number of features from the dataset. The algorithm then uses a majority vote to decide which splits are best and which prediction is the most accurate. To briefly describe bootstrap sampling, this process involves sampling with replacement from the dataset. Random forests are able to improve on predictions derived from decision trees; each tree fitted from each sample has a unique structure, meaning that the combination of all of these structures should more closely approximate the true data-generating function. The main tuning parameter in the random forest method is *mtry*, or the number of randomly selected variables used to create each tree. In addition to making more accurate predictions, random forest can be used to discern variable importance, which gives you a sense of the most crucial features in creating the trees and resulting predictions.

I intend to use the *train* function in the caret package in R in order to construct my set of trees for each position specified above. There are 53 features in the dataset; however, some of

them are linear combinations of the others. For example, there is data about passes, accurate passes, and percentage of passes that were accurate. I will try varying the number of features used in each tree, as well as the number of trees created by the function, in order to achieve the best prediction.

I will also utilize logistic regression to compare accuracy and significant effects on results in ACC soccer. Logistic regression is an algorithm that models the difference between two classes and attempts to predict the probabilities of each class. For each predictor in the dataset, the model creates a coefficient that relates the change in odds, once exponentiated, of predicting the positive class for a one-unit change in the predictor. These odds can be converted into probabilities through a simple mathematical transformation. The coefficients are evaluated with p-values showing significance for prediction at certain alpha levels. The model performance can be investigated through metrics such as accuracy, AIC, and kappa value. AUC is also widely used, especially in cases of response class imbalance. I utilize the *train* function in R with the "glm" method in order to create logistic regression models and use them to compare accuracy and significance of predictors from the random forest model.

Results

After running the random forest algorithm on the entire dataset and each of the position groups, I ran a logistic regression for the sake of comparing accuracy and variable significance of our models. To report results, I will first discuss the entire dataset and then describe each position group individually.

Entire Dataset

The random forest method produced an accuracy of .668 and a kappa value of .338 with 27 randomly selected variables. The most important variable by far was Playing in Scoring Attacks, which describes an action is part of an attack that ended up in a goal. Interceptions were regarded as the next most important predictor of results. The rest of the variables held similar importance metrics, but free balls picked up in the opposing half stood out slightly among that group.

Meanwhile, the logistic regression method produced an accuracy of .665 and a kappa value of 0.33. The model generated an AIC of 2385.8. The model included fourteen significant variables at the 0.05 alpha level. Most significant among them were Playing in Scoring Attacks, Grave Mistakes, and Interceptions. Playing in Scoring Attacks possessed a coefficient of 1.54 and Interceptions possessed a coefficient of 0.11, meaning that a higher number of these statistics led to a greater prediction of winning the game. Grave Mistakes, meanwhile, possessed a coefficient of -0.52, meaning that more grave mistakes committed would lead to a lower prediction of winning the game.

Center Backs

We now discuss the results for center backs. As noted above, this position is responsible for protecting the penalty area, the most dangerous zone of the pitch, while also starting attacks from deep. The random forest method produced an accuracy of .701 and a kappa value of .402 with 53 randomly selected predictors. Based on the variable importance metric, Grave Mistakes and Fouls had the biggest impact on classification. Tackles and Percentage of Accurate Passes also stood out, although they lagged significantly behind the former group.

The logistic regression method produced an accuracy of 0.663 with a kappa value of .327; it appears the random forest algorithm performed slightly better for center backs. The model generated an AIC of 494.2. This model also returned fourteen significant predictors at the 0.05 alpha level with Grave Mistakes and Fouls having significantly lower p-values than the rest. Looking at their coefficients, Grave Mistakes possessed an estimate of -0.84 and Fouls possessed an estimate of -0.74, showing that more mistakes and fouls would lead to a lower prediction of winning the game.

Defensive Midfielders

We now turn to defensive midfielders, whose main duties include protecting the defensive line, winning the ball back from the opposition midfielders, and distributing the ball to more advanced players. The random forest method produced an accuracy of 0.619 with a kappa value of 0.238 with 27 randomly selected predictors. The most important predictor for the role was Interceptions, followed by a few more predictors with a score greater than 70, including Lost Balls, Picking Up Free Balls in the Opposition Half, and Number of Passes Forward.

The logistic regression method produced an accuracy of 0.606 with a kappa value of 0.214. The model generated an AIC of 593.96. The method reported 9 significant variables at the 0.05 alpha level. Interestingly, the most significant variables with logistic regression were deemed to be Grave Mistakes, Interceptions, and Playing in Scoring Attacks. Extra Attacking and Key Passes and Yellow Cards were also significant, neither of which were included in the 20 most important variables for the random forest classification. Interceptions, Playing in Scoring Attacks, and Extra Attacking and Key Passes all had positive coefficients; Playing in Scoring

Attacks had the highest value at 1.58. Meanwhile, Grave Mistakes and Yellow Cards showed negative coefficients with Grave Mistakes having a higher absolute value.

Wingers

Wingers are the players who stretch the field and create space for the central players; they also get the ball in wide areas, drive at defenders, and create goal-scoring opportunities. The random forest method produced an accuracy of 0.655 and a kappa value of 0.31 with 2 randomly selected variables. Playing in Scoring Attacks was by far the most important variable; it had a variable importance score of 100, and the next closest was Accurate Passes Backward with 25.82. This model places importance on a few more passing stats, such as Passes Forward, Passes Backward, and Crosses.

The logistic regression method produces an accuracy of 0.607 and a kappa value of 0.213; it appears that the random forest method is slightly better for predicting impact of wingers. This model generates an AIC of 274.09, which is the best yet. The model includes only 4 significant predictors at the 0.05 alpha level: InStat Index, Playing in Scoring Attacks, Passes Accurate, and Percentage of Passes Accurate. Interestingly, while Passes Accurate has a positive coefficient, Percentage of Passes Accurate has a negative coefficient, albeit smaller in absolute terms.

Strikers

The striker is on the field to score goals, and his other, less important duties include holding up play and getting the rest of the team involved while acting as the first line of defense. The random forest method produces an accuracy of 0.671 with a kappa value of 0.344 with 53 randomly selected variables. Playing in Scoring Attacks is again the runaway leader in variable

importance with a value of 100. The next most important variables, Lost Balls, Dribbles, and Air Challenges won, register variable importance scores in the mid-teens.

The logistic regression method produces an accuracy of 0.637 with a kappa value of 0.273. It generates an AIC of 339.84. The model possesses 7 significant variables at the 0.05 alpha level with Playing in Scoring Attacks, Lost Balls, and Accurate Passes Forward representing the predictors with the lowest p-values. Playing in Scoring Attacks has a positive coefficient of 1.984, the highest of the relevant predictors in our analysis. Lost Balls also possesses a positive coefficient, but Accurate Passes Forward actually represents a negative coefficient of -0.43, which is not insignificant.

Discussion

From our analysis, we can discuss some insights gained about the nature of ACC soccer. First, it is important to establish the predictors that commonly appeared throughout the various sections of the procedure. Playing in Scoring Attacks constituted the single most important factor in predicting results across the entire analysis. This statistic is essentially an analog for Goals and Assists, so it makes sense that it would stand out. The percentage of accurate passes played by a player was also deemed significant by a variety of different models across different positions; this conclusion evidences another common tenet of soccer showing the importance of keeping possession and not turning the ball over to the opposition in dangerous areas.

Looking at the entire dataset, it appears that none of the positions themselves have a notably higher impact on result; each position dummy variable was deemed relatively insignificant by both algorithms. Both methods produced relatively almost identical accuracy and

kappa values, although the logistic regression generated a high AIC value. Playing in Scoring Attacks and Interceptions stood out in both types. We can see from the logistic regression coefficients and partial dependence plots that the probability of winning increases as both of these statistics increase. Interceptions, in particular, could provide an interesting insight into ACC Soccer; it shows the value of dispossessing the opponent, especially when they are passing the ball, and potentially taking that turnover and striking quickly on the counterattack. The idea that grave mistakes are harmful to winning is fairly obvious, but the metric of picking up free balls in the opponent's half also stood out among both algorithms. This could again point to the importance of picking up the ball quickly in dangerous areas and striking before the defense can get organized.

Focusing our discussion on the center backs, this position group actually created the random forest model that best described results in ACC soccer with an accuracy of around .7. Thus, we should probably emphasize looking at the variables deemed most important by that model. The two most important variables by far were Grave Mistakes and Fouls with scores of 100 and 97.54 respectively. Both partial dependence plots show steep negative slopes; it appears that accruing even two grave mistakes or fouls as a center back can have a very negative impact on your probability of winning the match. Tackles and Percentage of Accurate Passes also stood out in the random forest model, although these partial dependence plots showed much more interesting shapes. It actually appears that the more tackles that a center back makes, the lower your chances of winning the game. This relationship could represent the idea that your chances of winning get worse when the ball is in your half more often. It could also hint at the theory that pulling full backs out of position and forcing center backs into one-on-one situations where they

have to make tackles is a good strategy for winning games. Meanwhile, the partial dependence plots about the percentage of passes that were accurate follows a wave pattern after the 60% mark; chances of winning go up until around 68% before decreasing steadily until around 87% and increasing rapidly from there. I thought some of this variation could be due to lower frequencies of data describing the extremes of the statistic; thus, I created a histogram to check this hypothesis, and there actually exists a good amount of data from 65% to 95%. It seems that lower pass accuracy from center backs could actually point to a higher likelihood of winning the game. This relationship could point to the efficacy of long balls over the top of the opposition defense from center backs; these passes are more likely to be intercepted, but also bypass the clogged midfield and place more pressure on the opposing defense to deal with them, potentially leading to more mistakes and goal-scoring opportunities for your team. The logistic regression agreed with many of the above conclusions and also gave a negative coefficient of around -0.098 in the model, supporting the idea that a higher percentage of passes completed would lead to a lower chance of winning.

Turning to defensive midfielders, their models actually performed the worst out of the position groups in the analysis with accuracy scores of around .6. This could be attributed to the diversity of their role; defensive midfielders have a lot on their plate, and there are a lot of factors that contribute to winning soccer. Each algorithm emphasized different variables in their predictions; the random forest method highlighted Interceptions, Picking Up Free Balls in the Opp Half, Lost Balls, and Passes Forward while the logistic regression picked out Grave Mistakes, Playing in Scoring Attacks, and Extra Attacking and Key Passes as significant. Interceptions and tackles were the two common predictors among both models; however, both

models showed that a higher number of interceptions and lower number of tackles is conducive to wins in the ACC. These are both methods of dispossessing the opponent, but interceptions are so much more vital because they allow a team to earn clear possession and strike quickly while the defense is disorganized. Due to the relatively small difference in importance between many variables in the random forest model, I investigated a few more commonalities between this list and the significant variables at the .05 alpha level in the logistic regression model. The number of extra attacking and key passes, defined by InStat as a pass "creating 'a pre-goal-scoring moment' and leaving behind 3 or more opponent's players", positively affects winning as expected. However, it appears that a higher number of total attempted passes has a negative correlation with winning while a higher number of accurate passes has a positive correlation. This could point to the idea that defensive midfielders are supposed to prioritize safety in possession.

Now, we look at wingers at their results. The random forest algorithm was again slightly better at classification than the logistic regression algorithm; the former produced an accuracy of 0.655 and the latter produced an accuracy of .607. As we move further up the field, we begin to see the models emphasizing Playing in Scoring Attacks more and more. In this model, it has a variable importance score of 100 while the next variable in the hierarchy comes in at 25.82. Interestingly, the random forest model for wingers emphasizes passing stats more than any other position analyzed; Passes Forward, Passes Backward, Accurate Passes Backward, and Crosses were a few of the next most significant variables. Looking at partial dependence plots, wingers did not see the same sinusoidal pattern in percentage of passes that were accurate; this plot actually showed a relatively linear positive relationship from around 60% onwards. However, it

did show the same phenomenon regarding the different impacts of total attempted passes and accurate passes for results in ACC play. Although the logistic regression showed the lowest AIC of all positions, it only included 4 significant variables; this probably displays the sheer importance of goals and assists for the wing position in predicting results. This idea is reinforced by the fact that the best random forest model used only 2 predictors to classify observations. As players who mostly stay in wider areas of the pitch, they are involved in less actions than more central players, meaning their goal-scoring contributions are more highly emphasized in their roles.

Lastly, we analyze the results for strikers. Again, the runaway most important variable in both models was Playing in Scoring Attacks, which we acknowledge to focus our attention on more interesting occurrences. Lost Balls and Accurate Passes to the Left were deemed significant by both model types. Both the partial dependence plot and the regression coefficient showed a positive impact between the number of lost balls and winning games in the ACC. This might seem counterintuitive, but I feel that this shows how important it is that strikers get touches on the ball throughout the game, as it means they are involved and dangerous in the attack. The significance of accurate passes to the left hand side is also really interesting; I think it could point to the idea that most ACC teams tend to attack from the left-hand side in the final third. The variable also has a coefficient of 0.29 in the logistic regression model, which is relatively high for such a seemingly innocuous statistic. The logistic regression model actually highlighted passing stats; 4 of the 7 significant predictors had to do with passing. In addition to passing to the left, forward passing by strikers was seen to be a significant predictor of results in ACC soccer.

We can see a few patterns about the nature of ACC play coming up in our analysis. It seems that relatively obvious statistics such as Playing in Scoring Attacks and Grave Mistakes definitely do have a large impact on results. Specifically discussing attacking play, it seems that long balls over the top from center backs are favored more as a method of building up compared to short passing to more closely positioned players. A higher occurrence of these long balls could lead to more free balls as teams fight for possession, which would lead to more instances of defensive midfielders potentially picking up free balls in or close to the opposition half and establishing secure possession relatively high up the field. It appears that the models prioritized safety in possession from defensive midfielders, although the logistic regression picked up key passes as a significant predictor in regards to winning. This could hint at more creative play from deeper-lying playmakers having a positive effect, although it should be noted that both methods struggled to accurately predict results with statistics from defensive midfielders. Goals and assists are the defining factors for wingers, although passing, and pass accuracy specifically, were more relevant for wide players than any of the other position groups studied. Wingers generally have the least varied and most focused roles on the field; their job is to receive the ball from central players, make something happen, and create positive attacking passages. The analysis for strikers placed an even greater impact on playing in scoring attacks while also emphasizing lost balls and passing in the forward and left direction. As discussed above, the positive effect of lost balls is extremely intriguing because it shows the importance of keeping the game in the opposition half and giving strikers lots of touches to stay involved and try to make things happen. We could also theorize that most teams in the ACC direct their play down the left side of the field versus the right; this is common in soccer in general , as many of the

most dangerous players in the world, such as Neymar, Eden Hazard, and Sadio Mane, operate from the left before cutting in on their right foot.

Defensively, we can see that set pieces are extremely important in the ACC, a conclusion backed up by my own observations watching the games. Fouls by center backs have a very negative contribution towards winning games as they are usually committed in very dangerous areas of the field. As noted above, tackles from center backs also have a negative impact on results, perhaps because a center back is the last line of defense, and if they are attempting and winning more tackles, the other team is isolating them in dangerous positions. Defensive midfielders have a variety of different significant defensive duties, such as interceptions, picking up free balls, and tackles. Focusing on interceptions, it is significant that interceptions had a positive coefficient towards winning while tackles had a negative coefficient. I believe this speaks to the importance of counter-attacking in the ACC. When a pass is intercepted, the opposing player often establishes clear possession of the ball. Also, when a player passes the ball, the rest of their team moves in a manner that says they expect that pass to be completed and want to position themselves for the next pass in the sequence. Thus, when a pass is intercepted, the opposition has possession going forward against a disorganized defense. This allows them to move the ball forward quickly and with relatively little resistance. In a similar manner, picking up free balls is relevant because it gives one team possession at the expense of an opportunity for the other team to have the ball.

Further Analysis

My paper, while rigorous and thorough, could likely be taken in a number of different directions to improve accuracy and target different conclusions. First, it would be interesting to

investigate goalkeepers, fullbacks, and more attacking central midfielders to test different significant statistics in their play and have a more holistic view of how all the positions interact with each other in ACC soccer. However, soccer is and has always been a very difficult sport to analyze, mostly due to the massive impact that even a small action can have on the rest of the players. For instance, soccer is played in a variety of different formations, most of which are only rough descriptions of a player's position on the field. I think more accurate conclusions could be drawn by splitting analysis into formations and looking into the roles positions play within these team structures. Similarly, different teams approach the game in different ways; one team can prioritize possession and short passing, another team can try to be more vertical in their play, and another team can want as little of the ball as possible in the hopes of defending well and counter-attacking. While relatively impossible to quantitatively differentiate different styles without very detailed data, I think it would be conceivable to use possession as an analog to differentiate teams that prefer to have the balls from teams that don't, and then conduct an analysis based on that framework. Going in a different direction, I think manipulating the data further could prove beneficial. For instance, I removed all players with under 30 minutes played on the advice of some other people I've worked with, but perhaps lowering that restriction to 20 or 25 could be appropriate and expand the dataset a bit more. Also, instead of working with statistics covering total passes attended and number of accurate passes, you could just use the percentages and see if that improves accuracy; I was personally interested in using both because I wanted to see the effect of pass attempts as well, but I'm not convinced at the reliability of my conclusions in that regard. Additionally, getting more seasons of data could just give you a bigger sample size to work with.

Works Cited

Goller, D., Knaus, M. C., Lechner, M., & Okasa, G. (2018). Predicting Match Outcomes in Football by an Ordered Forest Estimator. *Economics Working Paper Series*, *1811*.

Schauberger, G., & Groll, A. (2018). Predicting matches in international football tournaments with random forests. *Statistical Modelling*, *18*(5-6), 460–482. doi: 10.1177/1471082x18799934

Ulmer, B., & Fernandez, M. (2014). Predicting Soccer Match Results in the English Premier League. Retrieved November 6, 2018, from http://cs229.stanford.edu/proj2014/Ben Ulmer, Matt Fernandez, Predicting Soccer Results in the English Premier League.pdf.