Match Officials and their Effect on European Football

Kaya Celebi

i. Introduction

It is evident that football (soccer) match officials' decisions are influential in the outcome of matches. Match officials (referees) determine whether to award fouls, penalties, set-pieces and whether to penalize with yellow or red cards. They play the key role of maintaining the official rules of the game while also making sure fair play and respectful behavior is retained. Many of the rules in football are defined with objective measures – ball out-of-bounds, goal scored, offside, etc. – however, there are also rules that require subjective decision-making by the referees[1]. With few, vague guidelines, it is up to the referees to make decisions about the harshness of a foul, the intent of a tackle, and the punishment for a player to keep them in line with fair play. There are general frameworks for these subjective decisions, however, they often vary from country-to-country which can cause controversy during inter-league and international competitions[2].

The application of these rules has some effect on the result of a football match. The key question of interest aims to answer by how much and how often match results are affected by these rulings. This question affects teams, FIFA/UEFA officials, and sports-betting: Teams could understand how they can alter tactics to make up for referee influence, FIFA/UEFA officials could implement new technologies/methods to overcome significant influence by referees, and sports-betters could learn to implement referee influence in their odds-making and predictions. Not to mention, fans would finally have some justification for being upset with a referee's decision during the weekly pub debates.

There are significant financial incentives for teams that would motivate discussion about this topic. Competition titles and final positions in the domestic league provide a significant influx of money for teams[3]. Each team wins some prize money scaled by their final league

[1] "FIFA," FIFA, accessed December 17, 2022, https://www.fifa.com/, 57.

[2] Vijay Murali, "Top 22 Worst Refereeing Decisions in World Football History," Bleacher Report (Bleacher Report, October 3, 2017), https://bleacherreport.com/articles/915954-top-22-worst-refereeing-decisions-in-world-football-history.

[3] "How Much Prize Money Do the Premier League Winners Make? Full List of Position Prizes." Sporting News, May 22, 2022. https://www.sportingnews.com/us/soccer/news/premier-league-prize-money-champions-full-table/ukfu5xbuzeogqcpvg0plkhe0.

position. For example, the winner of the English Premier League earns £44 million, second place earns £41.8 million, third place earns £39.6 million, and so on[4]. The results of these final positions are often determined on the very last match of the season, which puts millions of pounds on the line for a single game. Referee influence on these decisive matches can put clubs in undeserved deficits on the order of millions.

The concerns of match-fixing by bribing referees are very real in the footballing world[5]. Bribing a match official to favor a particular side by being more forgiving to them and being more strict to the opponent can clearly shift the odds of a game. Such corruption scams are often run by organized-crime rings that are embedded within sports-betting, whose aim is to make a certain bet a guaranteed win. These operations are incredibly detrimental to the sports-betting industry and the fans who engage in it. Being able to detect statistically significant influence from match officials can stamp out these operations.

This topic is a familiar one to football fans and teams all around the world. Despite this, it has not been investigated thoroughly. FIFA, UEFA, and the many other governing bodies of football are extremely protective of their match officials. They often go out of their way to penalize players, teams, and managers for criticizing or even speaking out against referee decisions in games, regardless of whether the criticisms are valid or not[6]. This makes discussion of this topic as well as reform within the governing bodies nearly impossible. If the discussion is supported by facts and statistics from many angles, then the tide can change in favor of fairer officiating.

The resulting models from this research were designed to predict a home-team win or loss given only subjective decisions made by the referee in a match. Using carefully-tuned Random Forests, we can predict match outcomes with 86% accuracy on unseen data in the English Barclays Premier League and 86% in the Spanish La Liga Santander.

[4] "How Much Prize Money Do the Premier League Winners Make? Full List of Position Prizes." Sporting News, May 22, 2022. https://www.sportingnews.com/us/soccer/news/premier-league-prize-money-champions-full-table/ukfu5xbuzeogqcpvg0plkhe0.

[5] https://www.forbes.com/sites/steveprice/2018/06/08/world-cup-2018-ghana-referee-bribes-scandal-shows-fifas-flaws/?sh=77f9c2563d46

[6] https://bleacherreport.com/articles/2579131-jose-mourinho-fined-given-suspended-1-match-ban-by-fa-after-referee-comments#:~:text=Chelsea-,Jose%20Mourinho%20Fined%2C%20Given%20Suspended%201%2DMatch%20Ban,by%20FA%20After%20Referee%20Comments&text=Chelsea%20manager%20Jose%20Mourinho%20has,he%20branded%20a%20%22disgrace.%22

ii. Project Goals

This project aims to detect the influence on match officials' subjective decisions on match outcomes. We choose to investigate the subjective decisions because the objective decisions (ex. out-of-bounds, goal, corner, etc.) are easily measurable and the officials are supported by extremely advanced technology that leaves little room for error.

If a model that uses exclusively subjective referee decisions can detect some advantage in favor of a home or away team, then we can show with statistical significance that these decisions have an influence on football matches as well as how much effect. It should be noted that this project *does not* aim to detect whether a referee makes an incorrect decision. The reason why subjective decisions are highly contested is because, although there are some decisions that are more obvious than others, there is still a gray area as to whether a decision is correct or not for many cases. Since there are no objective measures for these decisions, one often cannot reasonably make a convincing argument about the correctness of a decision.

It should also be noted that a match official's influence can easily be construed with a team's behavior. A team that accrues many yellow and red cards in a game can either have a strict referee, an undisciplined team, or both. Finding the correct side of the line between these two conclusions involves observing these trends over many teams, years, and leagues to avoid classification based on a team's behavior.

iii. Data Sources and Collection

The main data source used is the European Soccer Database[7] on Kaggle. It is an open-source database which documents data from over 25,000 matches from 2008 to 2016 in the top-flight league of 11 countries. The data was scraped and compiled from various live-betting websites which contained fragmented information regarding match and team data. It is supported by an Open Database License, allowing for nearly unrestricted usage. Most importantly, it includes the match results as well as detailed match events (fouls, cards, penalties, etc.). For this research, we will investigate the English Barclays Premier League and the Spanish La Liga Santander because they both have the highest number of data entries.

[7] Hugo Mathien, "European Soccer Database," Kaggle, October 23, 2016, https://www.kaggle.com/datasets/hugomathien/soccer.

iv. Data Preprocessing of Predictors

The data cleaning process involves obtaining the subjective decisions that referees make for each match and transforming them appropriately for off-the-shelf models. The transformation of this data involves loading the *Match* SQL table and manipulating it to obtain relevant predictors as well as unravel nested XML statements. Each row in the *Match* table corresponds to a single football match between two teams, with information regarding the teams that competed, the match outcome, match events, and betting odds. Of the 11 leagues in the database, only the English Barclays Premier League and Spanish La Liga Santander had enough data points for our purposes (n = 2873 and n = 1618, respectively).

The predictors we obtained from this table are the match events, which were listed as XML statements that included each foul, card, and their reason awarded in a match by timestamp. This was unraveled to obtain the number of fouls and cards of each color awarded to the home or away team faceted by the reason given by the referee. Any matches that had missing foul or card data were omitted from the dataset. Each foul and card event had a reason provided by the referee, such as “diving” or “pushing”. There are 13 different reasons each for fouls and cards that can be provided. We created dummy variables for each level of reason provided, and also faceted by whether the home or away team committed the foul/received the card. This yields 13 ✕ 2 ✕ 2 = 52 features from 13 reasons for fouls and cards (✕ 2) awarded to the home and away team (✕ 2). For example, the variable *away_team_foul_reason_from_behind = 2* describes for a single match, the away team committed two fouls where the player engaged from behind (Appendix Figure 1). In addition to these features, we included the number of home and away team yellow, red, and second-yellow cards obtained. This adds 2 ✕ 3 = 6 features, yielding a feature space in $\Re^{58}$.

After obtaining the finalized feature space, we observed whether any of the features were correlated with one another (see Appendix Figure 2). From the pairwise correlations, we observe that there are 4 pairs of features that have moderate correlation with one another (see Appendix Figure 3). From each pair, we can see that the features include the same reason for a card/foul but are for the home or away team. This implies that behaviors of pushing, tripping, and tackling from behind are mirrored in a moderate linear relationship between both teams, almost as “revenge” fouls. Although these features are not highly correlated, it is important to note them and remove them from the feature space, yielding 54 features. Additionally, 5 features that documented rare events (ex. goalkeeper handball) were padded with zeros, so they were also dropped, yielding 49 features.

v. Data Preprocessing of Response

The response variable we want to predict is the match result, which we encoded as “W”, “L”, or “D” for as a home win, home loss, or draw. There is an imbalance between the three classes (see Appendix Figure 4), which can be treated with downsampling and upweighting. For both the English Barclays Premier League (EPL) and the Spanish La Liga Santander (La Liga), the number of home-team draws is far too low for a multiclass classification problem (11.2% and 8.9%, respectively). Predicting only wins or losses and fixing the imbalance between those two classes is much more manageable, as there are only around twice as many losses as wins for both leagues. We can now proceed knowing that this is a binary classification problem.

Our imbalance can be solved by randomly downsampling the data from the larger class and upweighting that class by the factor that it is scaled down. For example, if a class is downsampled to have $\frac{1}{2}$ its initial size, it will be upweighted by a factor of 2. By downsampling the data, we now have both classes of equal size; and by upsampling the downsampled class, we are emphasizing its importance in the model given its prevalence in the original data. Using this methodology, we now have a class balance within the original data. We maintain this balance during the training process by stratifying the classes within the train-test split (ie. maintaining the initial class proportions in both the train and test folds).

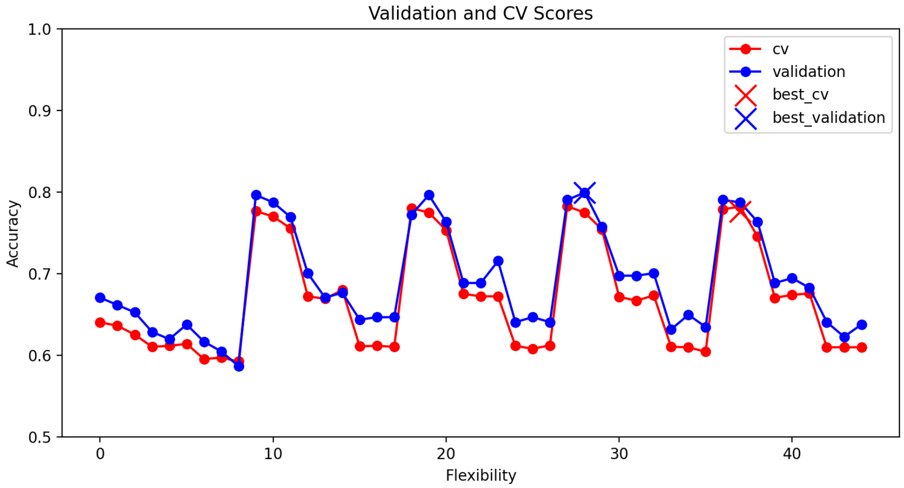
v. Exploratory Data Analysis

From preparing the data, it is clear that all of the variables we are working with are integers, and their distributions skew close to zero for many features. From the single-variable distributions of each feature (Appendix Figure 5), we can see that some are clustered at zero a few samples that have higher occurrence (ex. *home/away_team_card_reason_removing_shirt*) while a couple have a much larger range (ex. *home/away_team_card_color_y*). The distributions of nearly all of this data are consistent with reality (yellow cards are often frequently given out during all games, rarely are players ignorant enough to knowingly receive a card for removing their shirt). Given these distributions it is clear that any trends will be nonlinear and multidimensional. The feature space is much more limited when the data are integers (cardinality of the infinite sets are smaller), but having many features can make up for this with nonlinear patterns. When we stack the feature values for each match to observe their distribution as a whole dataset, we observe that each match is usually spread out along different subsets of features (Appendix Figure 6). This tells us that any patterns will not be mathematically simple to calculate, and most likely nonlinear.

This analysis shows us that classification through these decisions will most likely not be simple from a mathematical standpoint, but should be easily understandable from a heuristic standpoint. A nonlinear approach would be most appropriate, with interpretation and heuristics in mind. With that reasoning, a Random Forest Classifier would be the ideal model choice.

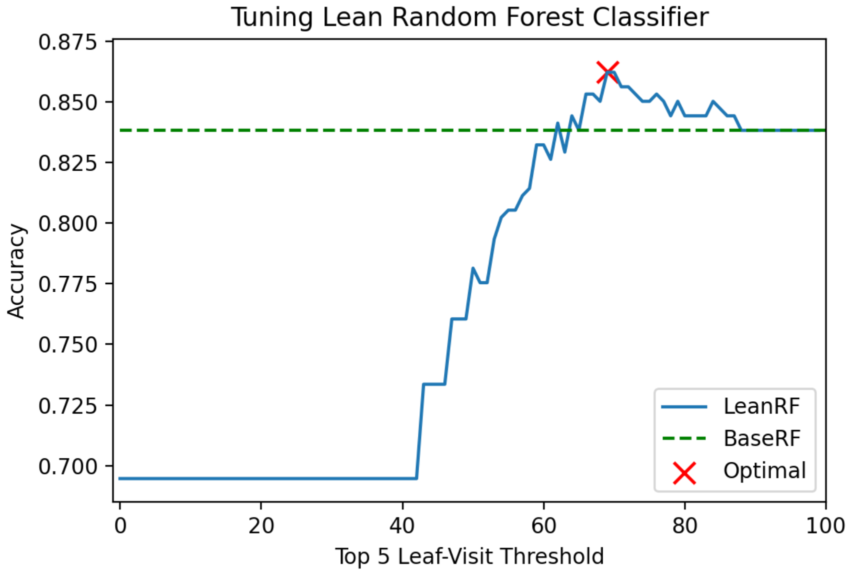
vi. Modeling with Random Forests

The process of designing the Random Forest involved applying the preprocessing pipeline to the data and tuning the model parameters with a grid search using cross-validation of *K*=10 folds. Our preprocessing handles the class imbalance that initially existed, allowing the model to weigh the data appropriately. We tuned over the maximum depth, minimum samples required for split, minimum samples required for leaf, and the minimum impurity decrease.



From the cross-validation tuning, we observe that our results underestimate the true validation score because, by reducing the training set by 10%, we are making the problem more difficult. We observe that the optimally-tuned classifier performed with training accuracies of 100% and validation accuracy of 84% on the EPL data (Appendix Figure 7) and 86% on the La Liga data (Appendix Figure 8).

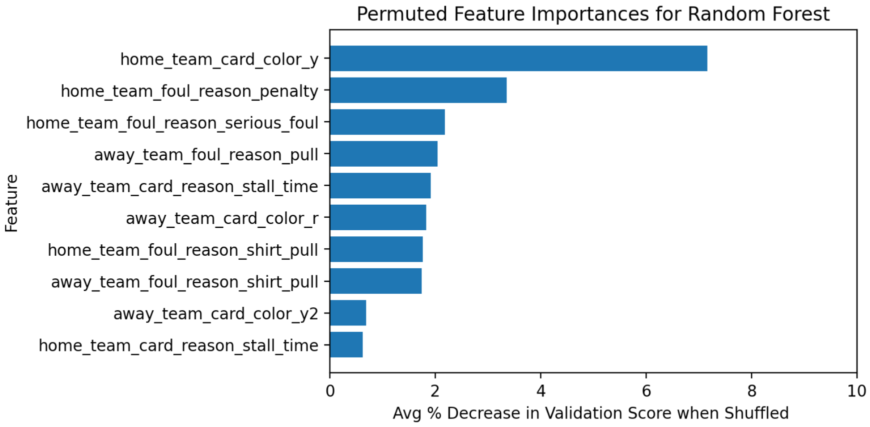
A further improvement on this model involved pruning the decision trees used in the classifier by retaining only the trees, whose leaf-nodes are sorted by number of visits, had at least 71% of samples visit the top t = 5 leaves. The purpose of this is to prioritize trees that had more applicable rules to the data and prune trees that would overfit to the noise. We obtained the threshold of 71% by tuning this “Lean Classifier” to obtain the best result.



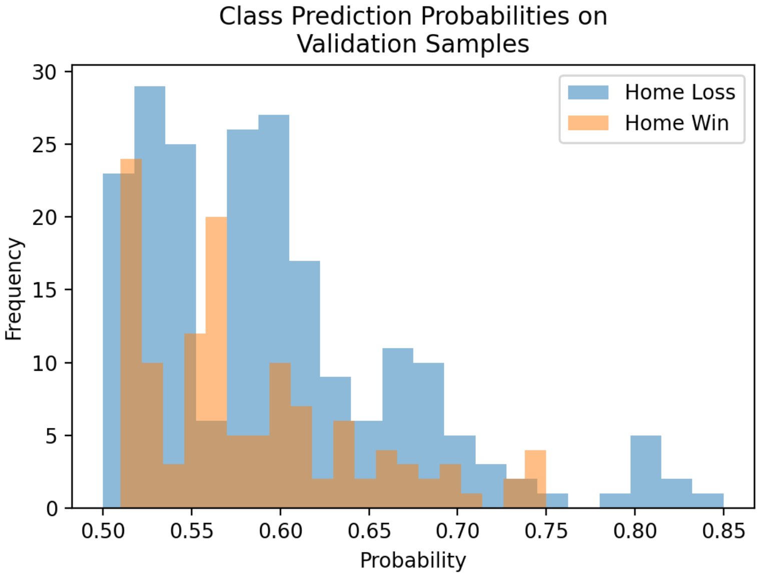
We observe that the Lean Classifier performed slightly better than the Random Forest on EPL data, obtaining a validation accuracy of 86% (Appendix Figure 9).

vii. Interpreting Results

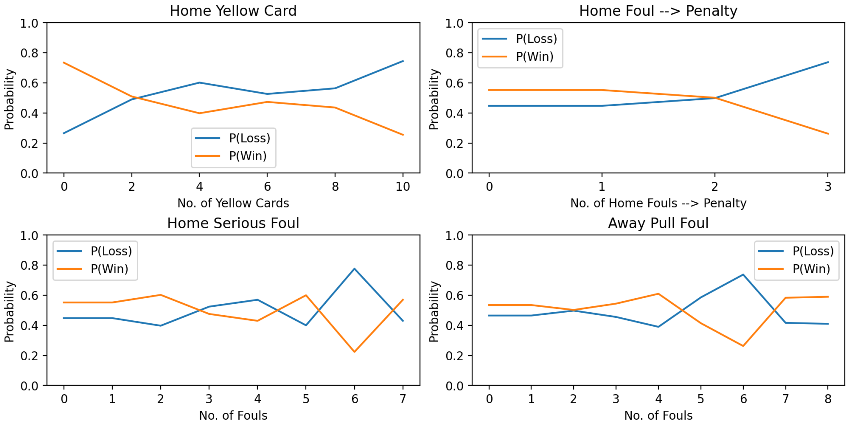
We can clearly determine from the model results that the subjective decisions made by a match official can be used to effectively predict the outcome of a game in the EPL and La Liga. It is now of interest to understand what the model chooses to prioritize when making these predictions.



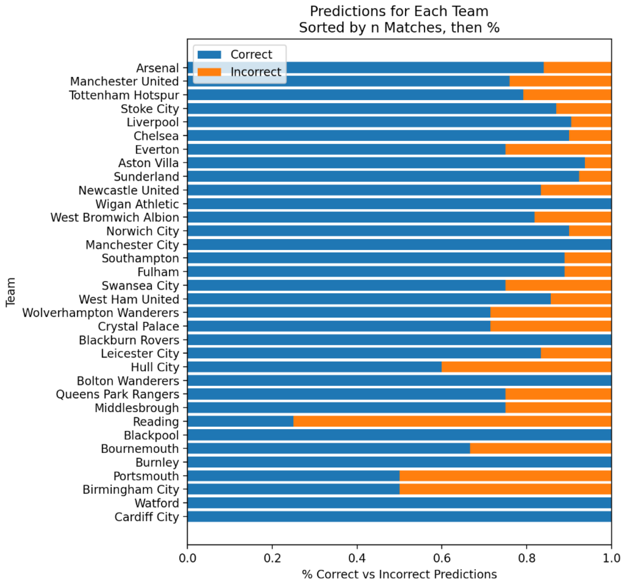
Permuted feature importances are calculated by shuffling the values in each feature and observing the resulting accuracy decrease in the validation data. We can perform these shufflings multiple times over each feature in order to obtain a consistent estimate of how important each feature is. We can then calculate which features are influential by obtaining those whose average score decrease is greater than zero by at least two times its standard deviation in score decrease. From the figure (above) we observe that the top three influential features relate to the number of yellow cards, penalties, and serious fouls that the home team receives. The total average percentage decrease for these significant features is 27.6%, which is a large margin of our validation accuracy about 50% (random chance).

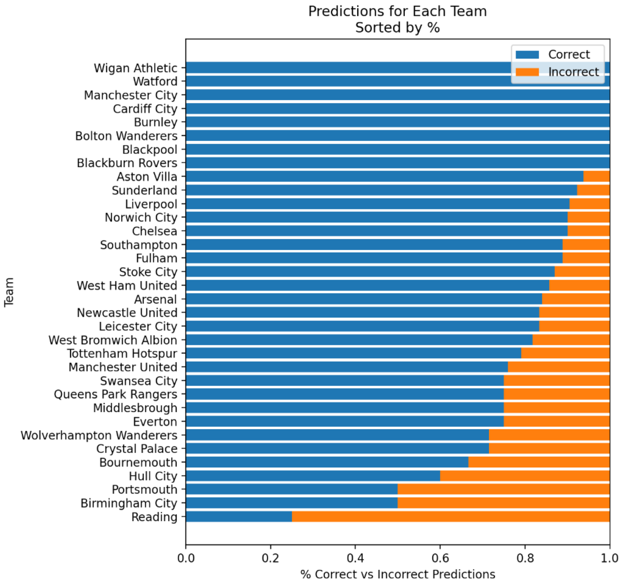


From the class prediction probabilities of each validation sample, we observe that there is quite a variety of certainty. It seems that when the model predicts a home loss, it is usually more uncertain, however, it has been more certain than when predicting a home win.



When we iterate these class probabilities faceted by certain features, we can get a better picture of how the model's certainty changes. From the top four features (above) we observe that the number of yellow cards and penalties that the home team concedes are useful for the model to improve its certainty. With fewer yellow cards, the home team is more likely to win, and as the number of yellow cards increases, this probability shifts in favor of a loss. This makes complete sense: these penalizations coincide with more opposition set pieces, tactical substitutions of players to avoid being sent off, players being sent off for a second yellow card, and player behavioral changes to avoid being penalized again that could lead to more lax defending. The number of penalties also makes sense: there is a high chance that the opposition can score from these. The number of serious fouls, though, remains uncertain until that number increases. The number of fouls by the away team when they pull a player also appears to be uncertain until that value increases.





In order to know that our model is attributing its predictions based on decisions and not behaviors of individual teams, it is important to consider how well the model predicts for individual teams. In the bar chart above, we can see the sorted correct/incorrect ratio for predictions made on the validation data on each Premier League team that appears in the dataset from 2008 to 2016 (total of 334 matches in validation set). When we sort by the number of predictions made, then by the correctness, we observe that the predictions aren't perfectly even. The model is able to predict many of the teams with 100% correctness. However, many of those teams are underrepresented in the validation, as seen in the plot on the left that sorts first by the number of times the team is predicted in the validation. When we take an average of the correctness (not a weighted average by number of appearances), it comes out to 84%, which is exactly our validation accuracy. This implies that our model is not learning from the behavior of the team, but rather the subjective decisions.

viii. Conclusion

From the Random Forest that was fitted over English Premier League and Spanish La Liga data, we are clearly able to see that the subjective decisions made by match officials are predictive of the match result. We can predict with 86% accuracy on unseen data, the match result as a win or loss in the Premier League and the Spanish La Liga. We found that, generally, the number of yellow cards, penalties, and serious fouls by the home team are the most significant predictors for understanding the match result (along with 7 others). We also found that the model is more confident and better at predicting home losses than wins with this data. Over thousands of matches, it is clear that the model is learning to make predictions based on these decisions rather than the overall behavior of each individual team. This is a significant result, in that, it sheds some light on a controversial topic within the football world. Managers, tacticians, and players can use these results to understand which facets of gameplay they should optimize in order to receive more favorable decisions from a referee.

This study was limited by the scarcity of data. Over eight years, there can be a total of {38 choose 2} × 8 = 5624 matches, of which many were missing the data that was required for this research. This also limited the scope of the study, as we were limited to studying only two leagues out of the eleven that were available. For the future, it would be interesting to see if individual referees have different influences on a match result, as well as further discussion on the interactions of these variables.

References

FIFA. Accessed December 17, 2022. https://www.fifa.com/.

"How Much Prize Money Do the Premier League Winners Make? Full List of Position Prizes." Sporting News, May 22, 2022. https://www.sportingnews.com/us/soccer/news/premier-league-prize-money-champions-full-table/ukfu5xbuzeogqcpvg0plkhe0.

Mathien, Hugo. "European Soccer Database." Kaggle, October 23, 2016. https://www.kaggle.com/datasets/hugomathien/soccer.

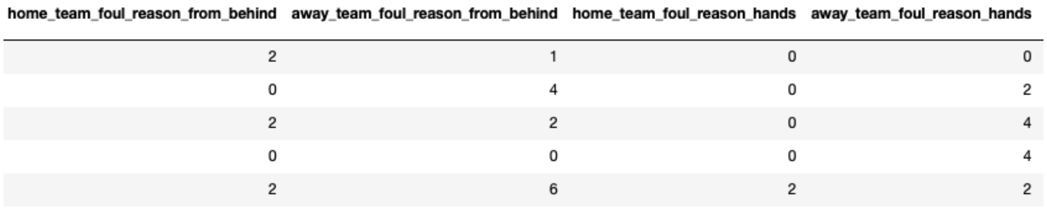
Murali, Vijay. "Top 22 Worst Refereeing Decisions in World Football History." Bleacher Report. Bleacher Report, October 3, 2017. https://bleacherreport.com/articles/915954-top-22-worst-refereeing-decisions-in-world-football-history.

Price, Steve. "World Cup 2018: Ghana Referee Bribes Scandal Shows Fifa's Flaws." Forbes. Forbes Magazine, June 8, 2018. https://www.forbes.com/sites/steveprice/2018/06/08/world-cup-2018-ghana-referee-bribes-scandal-shows-fifas-flaws/?sh=77f9c2563d46.

Sunderland, Tom. "Jose Mourinho Fined, given Suspended 1-Match Ban by FA after Referee Comments." Bleacher Report. Bleacher Report, September 24, 2017. https://bleacherreport.com/articles/2579131-jose-mourinho-fined-given-suspended-1-match-ban-by-fa-after-referee-comments#:~:text=Chelsea-,Jose%20Mourinho%20Fined%2C%20Given%20Suspended%201%2DMatch%20Ban,by%20FA%20After%20Referee%20Comments&text=Chelsea%20manager%20Jose%20Mourinho%20has,he%20branded%20a%20%22disgrace.%22.

Appendix

Figures and Tables



home_team_foul_reason_from_behind	away_team_foul_reason_from_behind	home_team_foul_reason_hands	away_team_foul_reason_hands
2	1	0	0
0	4	0	2
2	2	0	4
0	0	0	4
2	6	2	2

Figure 1. Sample image of data matrix displaying a few columns and their values.

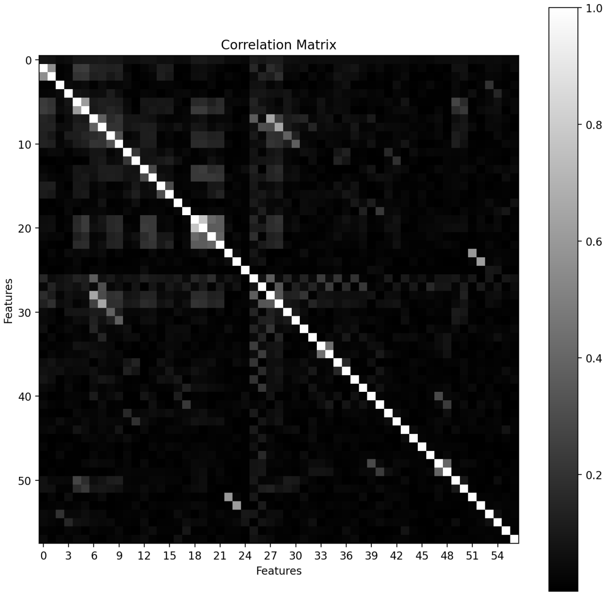
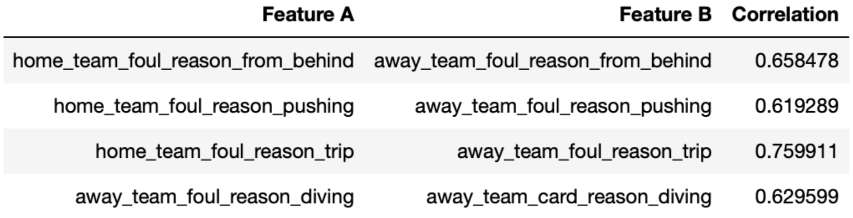
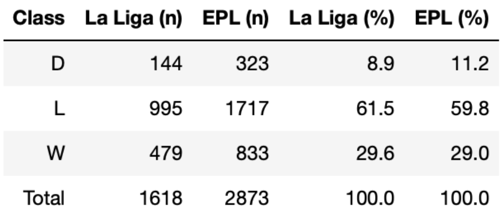


Figure 2. Correlation matrix heatmap for the initial 58 features. There are no features that are highly correlated (> 0.9), however, there are a few that are somewhat correlated. Dropping correlated columns yields a feature space in $\Re^{49}$.



Feature A	Feature B	Correlation
home_team_foul_reason_from_behind	away_team_foul_reason_from_behind	0.658478
home_team_foul_reason_pushing	away_team_foul_reason_pushing	0.619289
home_team_foul_reason_trip	away_team_foul_reason_trip	0.759911
away_team_foul_reason_diving	away_team_card_reason_diving	0.629599

Figure 3. Features with the highest pairwise correlation in the dataset. Only 4 pairs above correlation = 0.6.



Class	La Liga (n)	EPL (n)	La Liga (%)	EPL (%)
D	144	323	8.9	11.2
L	995	1717	61.5	59.8
W	479	833	29.6	29.0
Total	1618	2873	100.0	100.0

Figure 4. Summary statistics for the EPL and La Liga datasets. There is a clear class imbalance between home wins, draws, and losses.

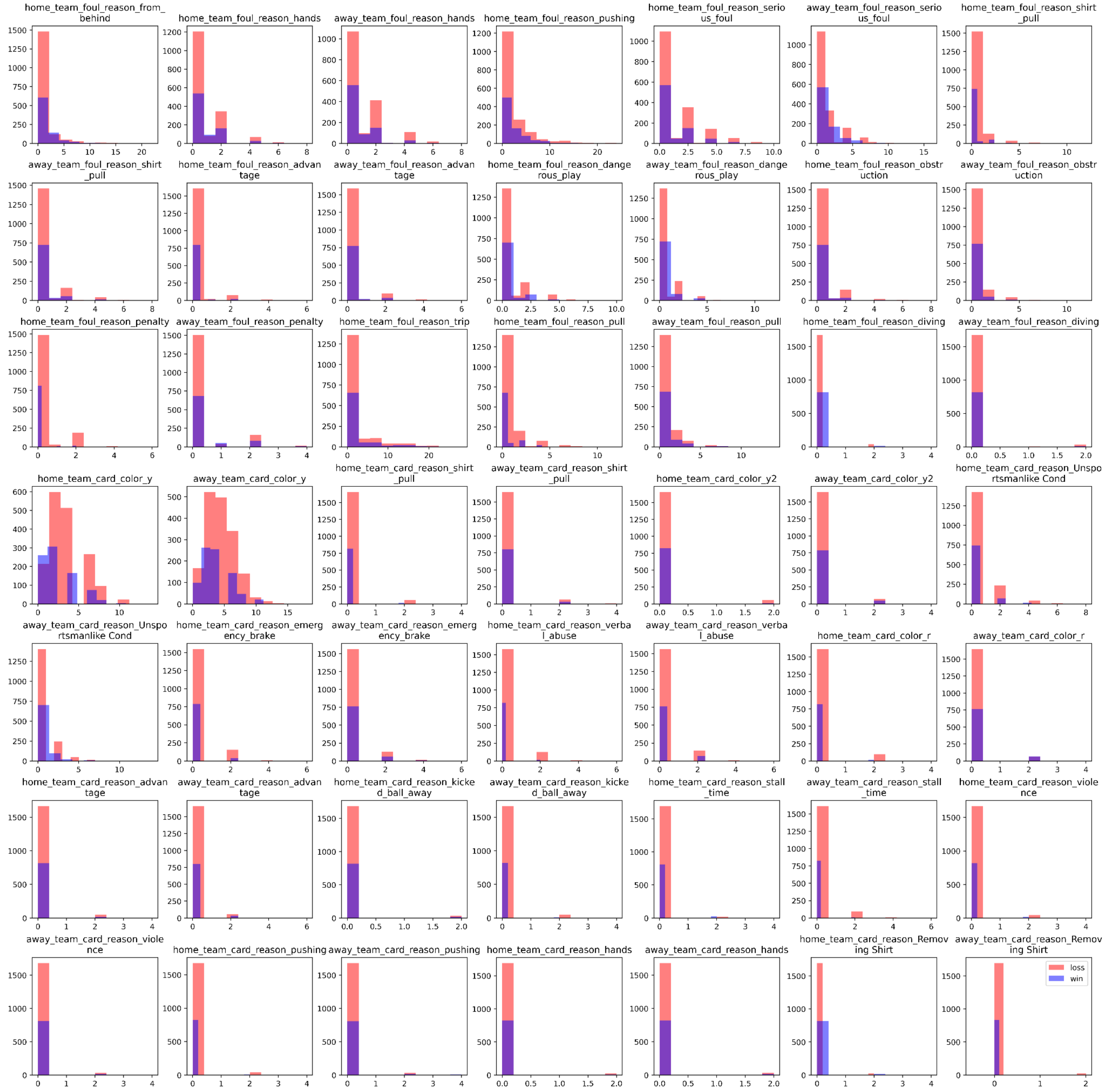


Figure 5. Single-variable distributions for EDA of our 49 predictors. We observe that many are clustered at 0, while a few predictors have a wider distribution.

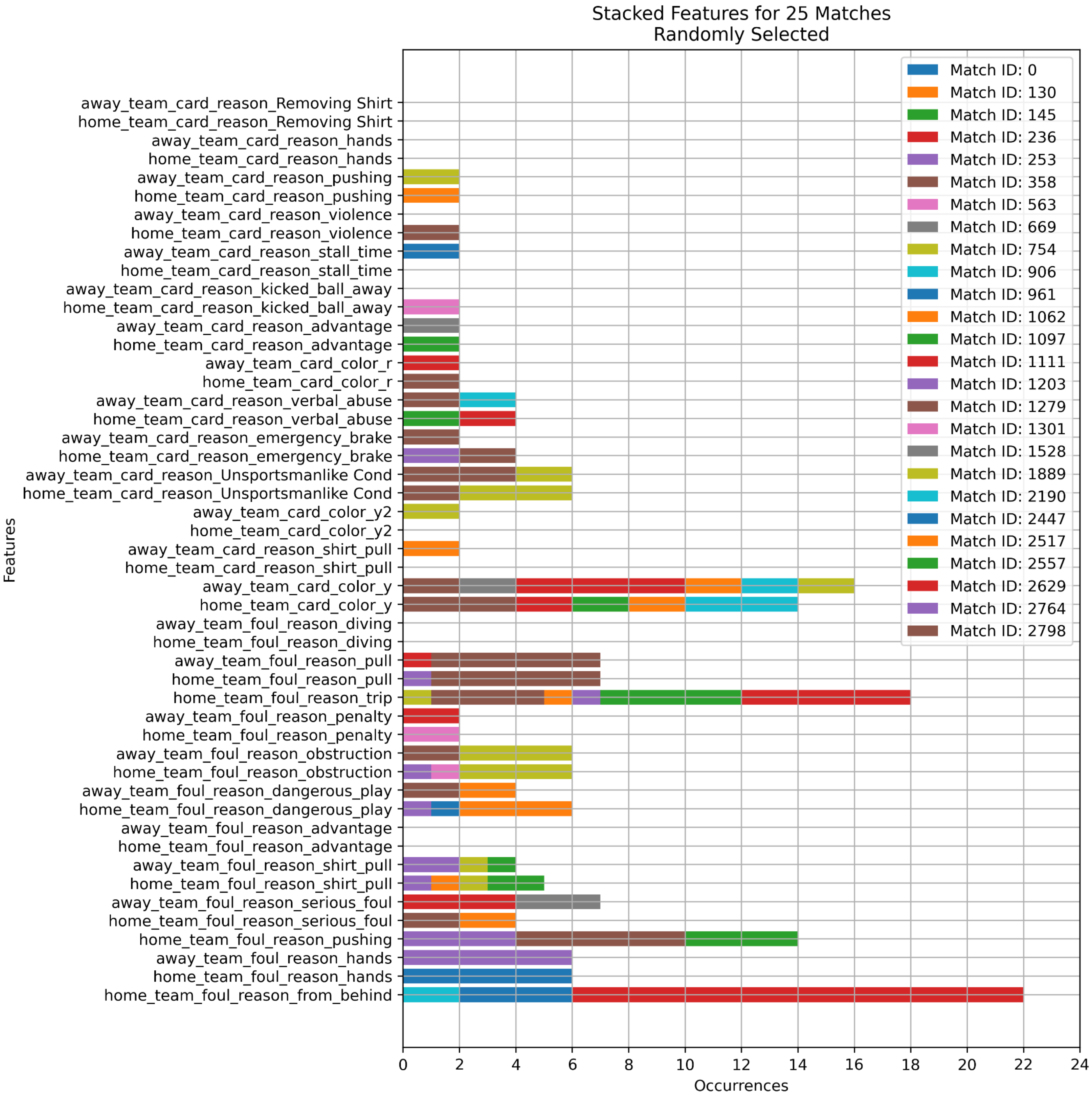
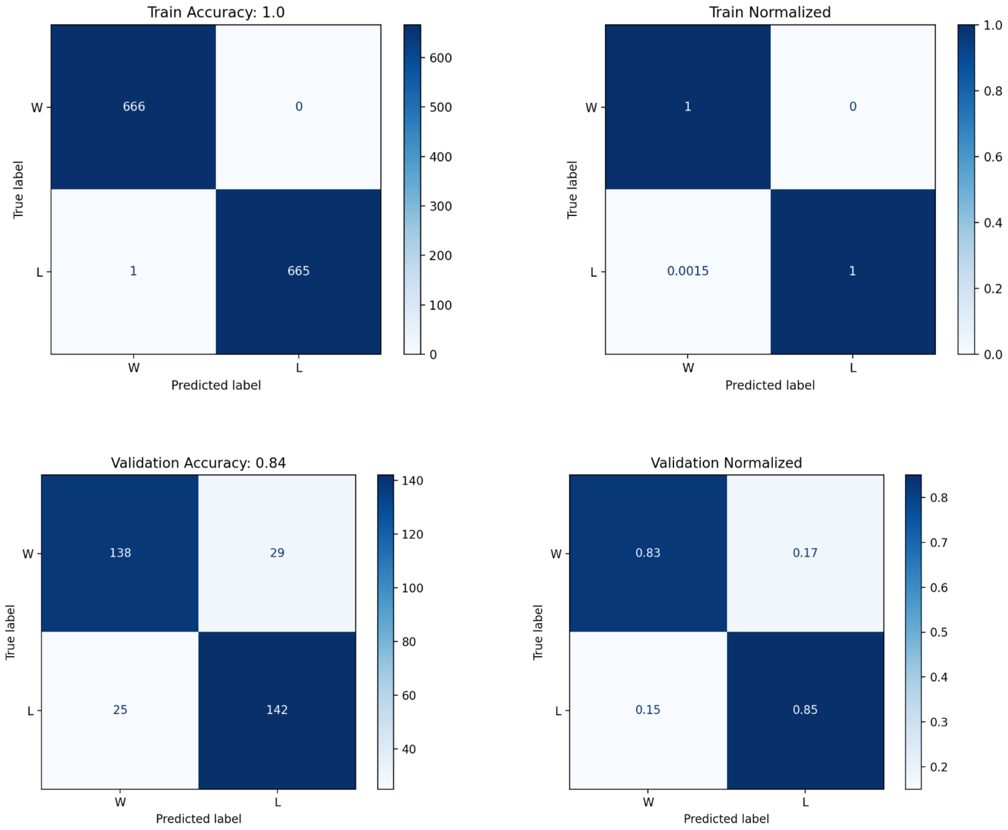
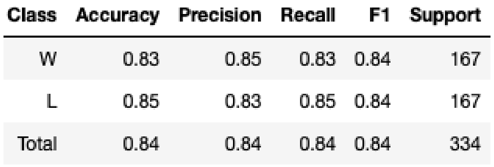


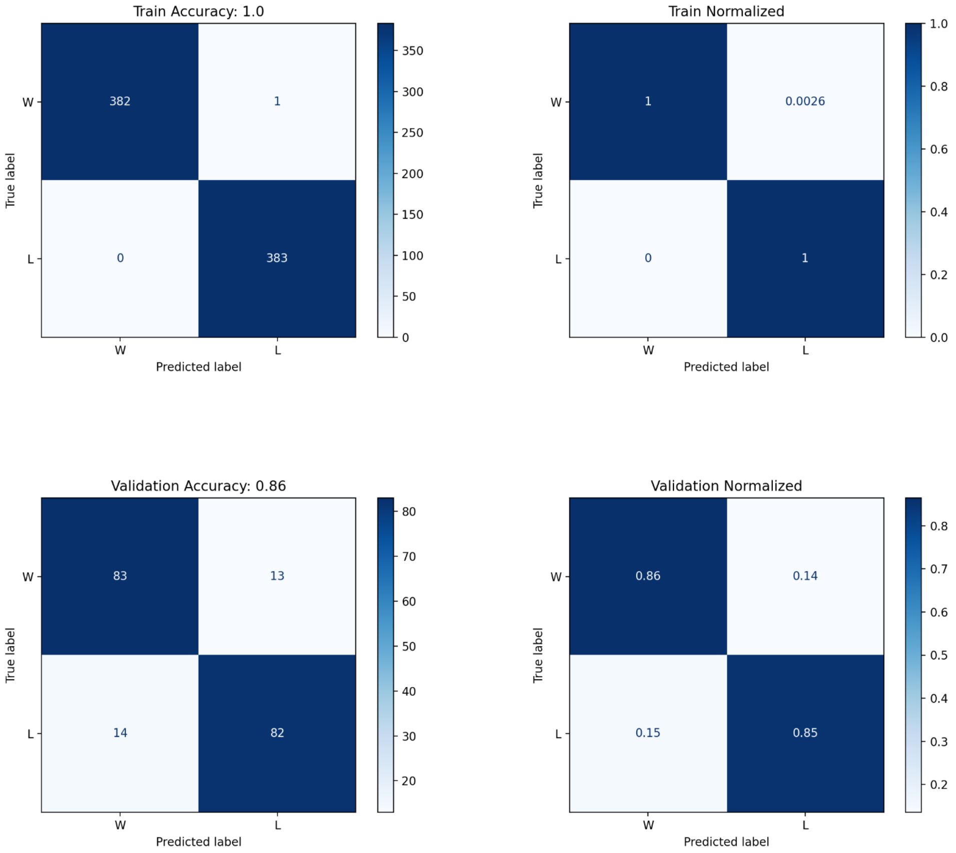
Figure 6. Stacked predictors for 25 randomly selected matches. We observe the values for each match are spread along different subsets of predictors, indicating that any pattern would be mathematically complex, but there could exist a fairly reasonable heuristic.

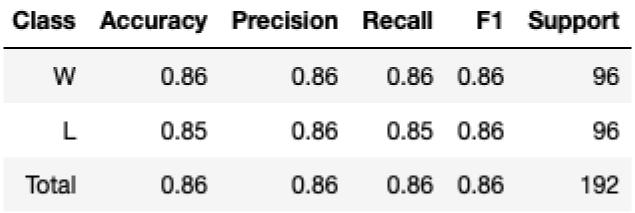




Class	Accuracy	Precision	Recall	F1	Support
W	0.83	0.85	0.83	0.84	167
L	0.85	0.83	0.85	0.84	167
Total	0.84	0.84	0.84	0.84	334

Figure 7. Training and validation confusion matrices for best hyper-parameter setup on EPL data. Received 100% training accuracy, 84% validation accuracy.





Class	Accuracy	Precision	Recall	F1	Support
W	0.86	0.86	0.86	0.86	96
L	0.85	0.86	0.85	0.86	96
Total	0.86	0.86	0.86	0.86	192

Figure 8. Training and validation confusion matrices for best hyper-parameter setup on La Liga data. Received 100% training accuracy, 86% validation accuracy.

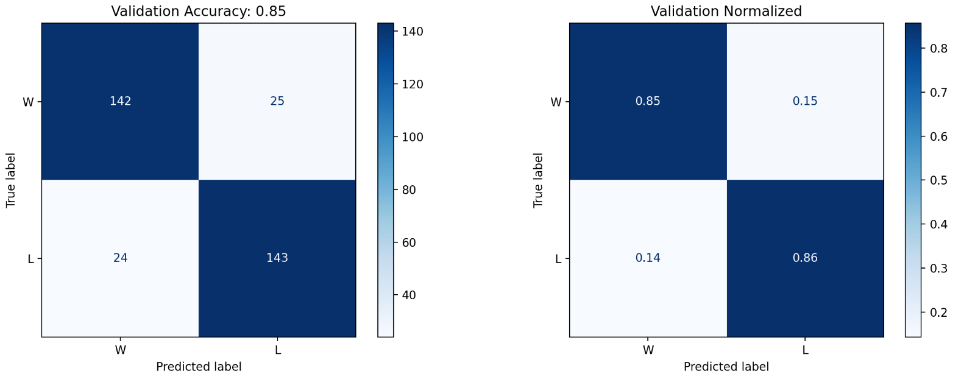


Figure 9. Validation confusion matrix for best tuning result for the “Lean Classifier.” Performs only slightly better than the off-the-shelf Random Forest.