Predicting European Football Results

By Ben Royles

Overview

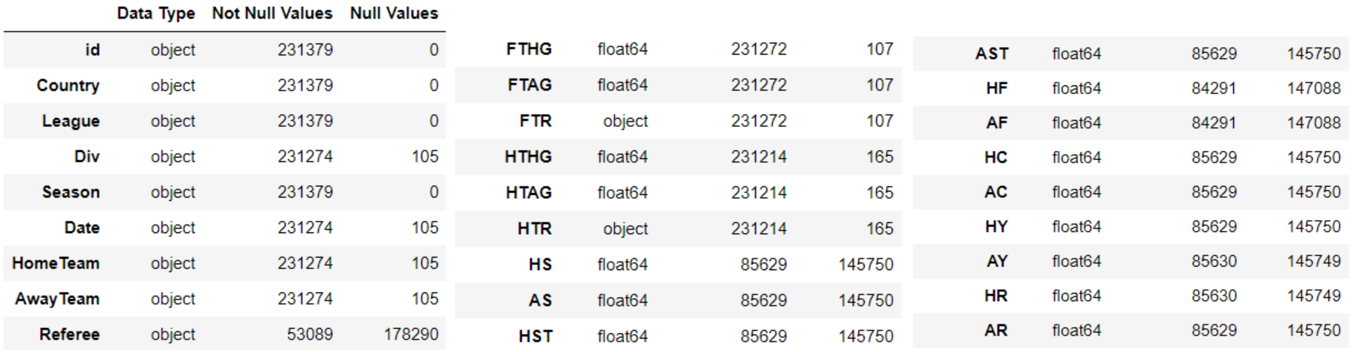
Can European Football results be predicted with a high degree of accuracy? If possible, there is value in a monetary sense as well as social value. Football is the market leader in the Sports betting and Fantasy sports markets, both of which are predicted to almost double in value by 2030. Socially, a successful model would give fans a greater understanding of the game and the factors that help determine the final outcome.

Background

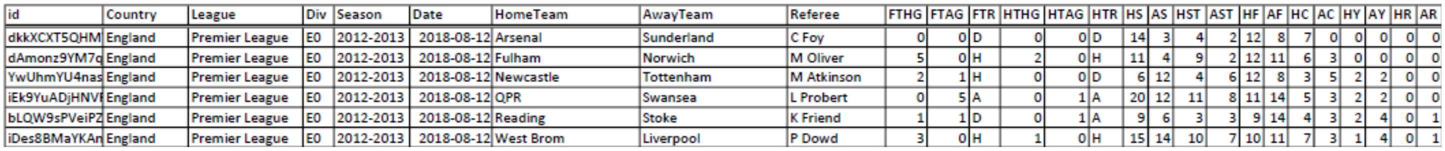
Football is typically a low scoring game, certainly when compared to the most popular North American sports. It is a game of fine margins, where the difference between winning, drawing or losing can come down to 1 or 2 key moments in the space of 90 minutes. This makes accurately predicting the outcome far more challenging and nuanced than in some other sports.

Dataset

The dataset was downloaded from https://football-data.co.uk . The initial format included team and league information, the final result and advanced statistics such as shots on target and fouls; a full glossary can be found in the appendix.



	Data Type	Not Null Values	Null Values
id	object	231379	0
Country	object	231379	0
League	object	231379	0
Div	object	231274	105
Season	object	231379	0
Date	object	231274	105
HomeTeam	object	231274	105
AwayTeam	object	231274	105
Referee	object	53089	178290
FTHG	float64	231272	107
FTAG	float64	231272	107
FTR	object	231272	107
HTHG	float64	231214	165
HTAG	float64	231214	165
HTR	object	231214	165
HS	float64	85629	145750
AS	float64	85629	145750
HST	float64	85629	145750
AST	float64	85629	145750
HF	float64	84291	147088
AF	float64	84291	147088
HC	float64	85629	145750
AC	float64	85629	145750
HY	float64	85629	145750
AY	float64	85630	145749
HR	float64	85630	145749
AR	float64	85629	145750

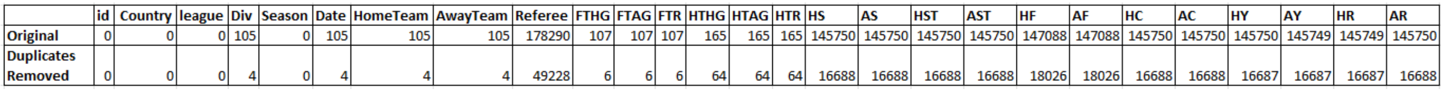


id	Country	League	Div	Season	Date	HomeTeam	AwayTeam	Referee	FTHG	FTAG	FTR	HTHG	HTAG	HTR	HS	AS	HST	AST	HF	AF	HC	AC	HY	AY	HR	AR
dkkXCXT5QHM	England	Premier League	E0	2012-2013	2018-08-12	Arsenal	Sunderland	C Foy	0	0	D	0	0	D	14	3	4	2	12	8	7	0	0	0	0	0
dAmonz9YM7q	England	Premier League	E0	2012-2013	2018-08-12	Fulham	Norwich	M Oliver	5	0	H	2	0	H	11	4	9	2	12	11	6	3	0	0	0	0
YwUhmYU4nas	England	Premier League	E0	2012-2013	2018-08-12	Newcastle	Tottenham	M Atkinson	2	1	H	0	0	D	6	12	4	6	12	8	3	5	2	2	0	0
iEk9YuADjHNVI	England	Premier League	E0	2012-2013	2018-08-12	QPR	Swansea	L Probert	0	5	A	0	1	A	20	12	11	8	11	14	5	3	2	2	0	0
bLQW9sPVeiPZ	England	Premier League	E0	2012-2013	2018-08-12	Reading	Stoke	K Friend	1	1	D	0	1	A	9	6	3	3	9	14	4	3	2	4	0	1
iDes8BMaYKAn	England	Premier League	E0	2012-2013	2018-08-12	West Brom	Liverpool	P Dowd	3	0	H	1	0	H	15	14	10	7	10	11	7	3	1	4	0	1

These statistics can explain the match with a good level of detail but obviously only exist once the match has occurred and would not be available to predict outcomes ahead of time

Cleaning and Preprocessing

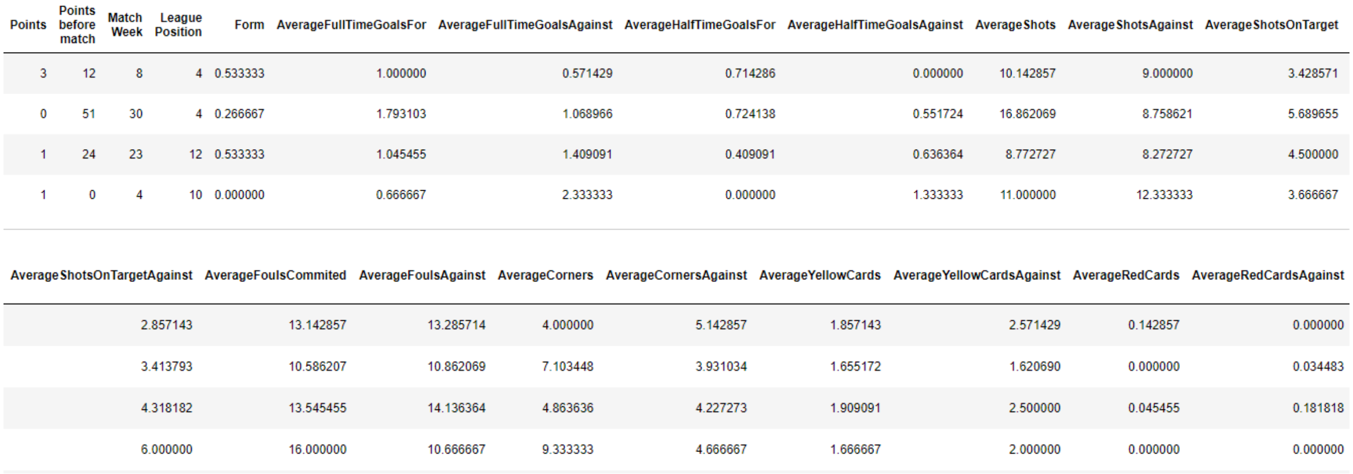
Similar to most datasets, the first challenge was to tackle duplicated and missing data. Almost every column had missing data that was spread throughout the leagues and seasons which posed a number of problems. The most significant was that simply deleting the data could dramatically impact season running statistics, impacting averages and league positions. I discovered through extensive exploratory data analysis (EDA) that a large proportion of the missing data was in the form of partial duplicates. The data and teams involved matched but other columns were nulls. Once these were removed, the number of actual null values was significantly reduced.



	id	Country	league	Div	Season	Date	HomeTeam	AwayTeam	Referee	FTHG	FTAG	FTR	HTHG	HTAG	HTR	HS	AS	HST	AST	HF	AF	HC	AC	HY	AY	HR	AR
Original	0	0	0	105	0	105	105	105	178290	107	107	107	165	165	165	145750	145750	145750	145750	147088	147088	145750	145750	145750	145749	145749	145750
Duplicates Removed	0	0	0	4	0	4	4	4	49228	6	6	6	64	64	64	16688	16688	16688	16688	18026	18026	16688	16688	16687	16687	16687	16688

I made the decision to remove the Referee column entirely as it was inconsistently populated to offer any value. Additionally, I don't believe that a specific referee has much impact on the outcome of the game. I was able to resolve the remaining null values, with the exception of the advanced statistics, through selective removal, interpolation and manually searching for the missing information.

The main challenge with this dataset was to transform it from descriptive to predictive. Firstly, I duplicated the dataframe and adjusted the column names so each match had an entry from both team's perspectives which enabled me to calculate league statistics such as Points, League Position and Form. Advanced statistics were also transformed into season averages. After these calculations, I shifted all of the values so that they explain what position the team was in and what their averages were prior to the match beginning instead of after it.

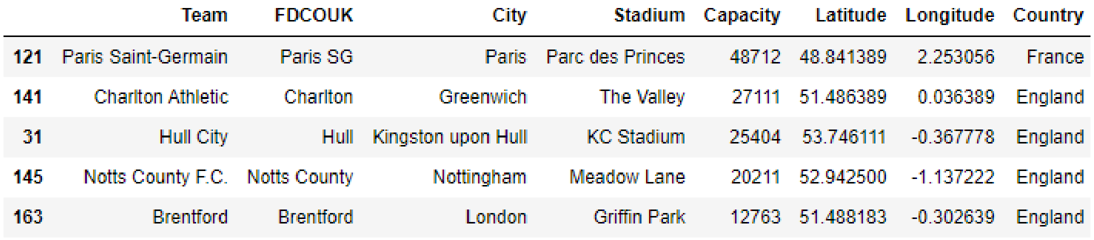


Points	Points before match	Match Week	League Position	Form	AverageFullTimeGoalsFor	AverageFullTimeGoalsAgainst	AverageHalfTimeGoalsFor	AverageHalfTimeGoalsAgainst	AverageShots	AverageShotsAgainst	AverageShotsOnTarget
3	12	8	4	0.533333	1.000000	0.571429	0.714286	0.000000	10.142857	9.000000	3.428571
0	51	30	4	0.266667	1.793103	1.068966	0.724138	0.551724	16.862069	8.758621	5.689655
1	24	23	12	0.533333	1.045455	1.409091	0.409091	0.636364	8.772727	8.272727	4.500000
1	0	4	10	0.000000	0.666667	2.333333	0.000000	1.333333	11.000000	12.333333	3.666667

AverageShotsOnTargetAgainst	AverageFoulsCommited	AverageFoulsAgainst	AverageCorners	AverageCornersAgainst	AverageYellowCards	AverageYellowCardsAgainst	AverageRedCards	AverageRedCardsAgainst
2.857143	13.142857	13.285714	4.000000	5.142857	1.857143	2.571429	0.142857	0.000000
3.413793	10.586207	10.862069	7.103448	3.931034	1.655172	1.620690	0.000000	0.034483
4.318182	13.545455	14.136364	4.863636	4.227273	1.909091	2.500000	0.045455	0.181818
6.000000	16.000000	10.666667	9.333333	4.666667	1.666667	2.000000	0.000000	0.000000

Once this was done, I transformed the data frame back into the original structure with the addition of each team's season running statistics prior to the match happening. The final data frame is similar to the one above, with each column appearing twice, once for the home team and once for the away team.

One thing that became clear from looking closely at the data is that home teams win more often than any other outcome. In order to gain more insight and hopefully provide more accurate predictions, I added stadium specific data to the data frame. This was scrapped from https://opisthokonta.net/?p=619



	Team	FDCOUK	City	Stadium	Capacity	Latitude	Longitude	Country
121	Paris Saint-Germain	Paris SG	Paris	Parc des Princes	48712	48.841389	2.253056	France
141	Charlton Athletic	Charlton	Greenwich	The Valley	27111	51.486389	0.036389	England
31	Hull City	Hull	Kingston upon Hull	KC Stadium	25404	53.746111	-0.367778	England
145	Notts County F.C.	Notts County	Nottingham	Meadow Lane	20211	52.942500	-1.137222	England
163	Brentford	Brentford	London	Griffin Park	12763	51.488183	-0.302639	England

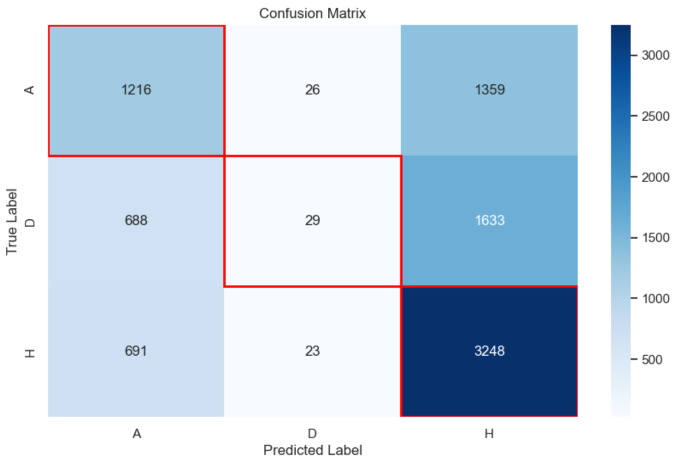
It included information about the geographic location in latitude and longitude, the team that plays there, and the capacity. Using this information, I calculated the distance the away team had to travel. I had hoped to find attendance data for every match but this was not possible, so for the purpose of this project, I have assumed that capacity and attendance are interchangeable and that each team utilizes the same proportion of their capacity for every game.

Insight and Results

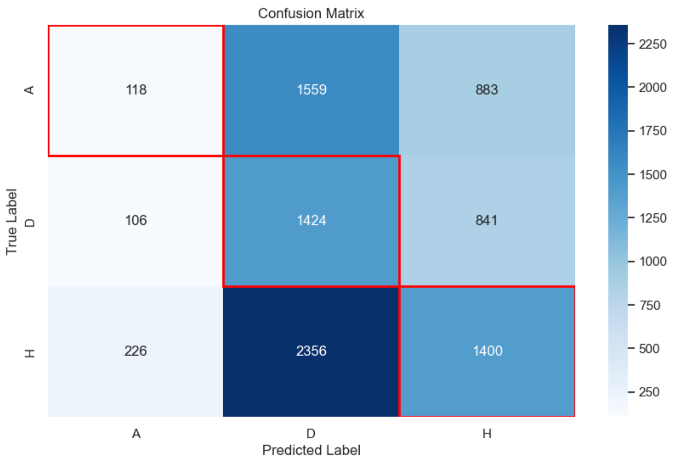
I initially created models using only the data that had stadium information, which left me with around 21000 rows. After some experimentation, I concluded that it was better to remove the stadium information and create models using the larger dataset.

I ran some basic models to get a baseline and produced confusion matrices to highlight where they performed well and where they have weaknesses. The highlighted, red diagonals are the correct predictions.

Logistic regression 44.6%



Decision Tree 46.2%



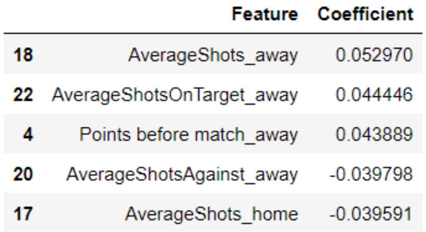
The two models have similar overall accuracies, however, each model had its own distinct weaknesses. Logistic Regression overpredicts home wins, while the Decision tree overpredicts draws and predicts almost no away wins.

I proceeded to build and tune a variety of models in the hope of producing one that performs better. The hyper parameters were tuned using accuracy but I have included a number of evaluation metrics in the analysis as the importance of any particular one would be determined by the use case. I elected to keep the actual features I created and not undertake principal component analysis or any other dimensionality reduction as I want my models to be interpretable. As well as this, I experimented with balancing the classes in the training data using undersampling but this actually reduced the overall accuracy. Additionally, I experimented with feature forward selection, this too resulted in lower accuracy.

Logistic Regression

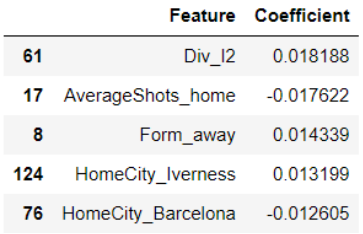
As a result of tuning hyperparameters I was able to score an accuracy of 50.3% on the test set which is quite a significant improvement. This was largely due to tuning the C parameter. Reducing the value of C and, therefore, increasing the strength of regularization, prevented the model from overfitting the training data. This resulted in better performance with the unseen test data. Looking at the top 5 coefficients for each class highlights the issue the model has predicting draws.

Away Win



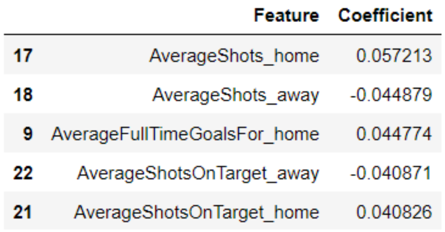
	Feature	Coefficient
18	AverageShots_away	0.052970
22	AverageShotsOnTarget_away	0.044446
4	Points before match_away	0.043889
20	AverageShotsAgainst_away	-0.039798
17	AverageShots_home	-0.039591

Draw



	Feature	Coefficient
61	Div_I2	0.018188
17	AverageShots_home	-0.017622
8	Form_away	0.014339
124	HomeCity_Iverness	0.013199
76	HomeCity_Barcelona	-0.012605

Home Win



	Feature	Coefficient
17	AverageShots_home	0.057213
18	AverageShots_away	-0.044879
9	AverageFullTimeGoalsFor_home	0.044774
22	AverageShotsOnTarget_away	-0.040871
21	AverageShotsOnTarget_home	0.040826

As you can see, away win and home win have much higher coefficients than draw, highlighting why the model predicts so few draws. Another interesting observation is that for away wins, the most important features are their own, whereas home win has a more even split of both their statistics and the opposing teams.

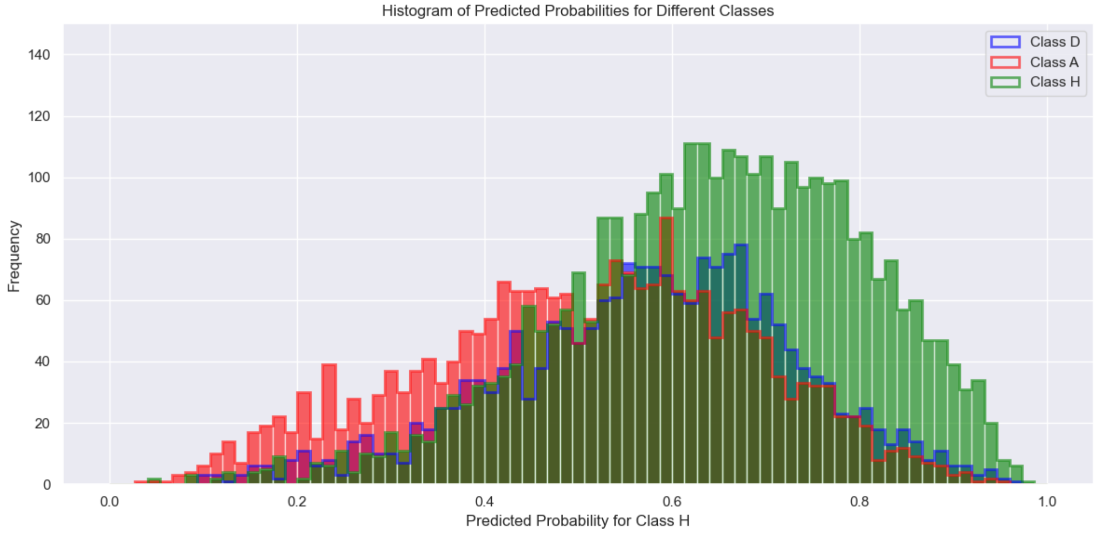
Decision Trees

I created a number of decision tree models, initially from tuning the hyper parameters and then advanced to random forest, gradient boot and XGboost models. All of these showed some improvement from the original decision tree model with accuracies of between 48-50% on the test set. The XGBoost model scored the best with an accuracy of 50.5%

Binary Classification

Since most models struggled to classify draws, I experimented with using a binary logistic regression classifier for classes A and H. This actually caused a slight increase in accuracy with the test set. However, as the validation set scored lower, this may just be due to chance.

The diagram below illustrates the binary classification model, with colors representing actual outcomes (A, D, and H) in the testing set. The histograms depict the probabilities assigned by the model for each class based on input features. Probabilities above 0.5 are predicted as class H, and those below are class A. While there is some bimodality between classes A and H, significant overlap exists among all three classes. This overlap contributes to the models misclassifying many observations, particularly class D, which is likely to apply to all models.



I also manually set the decision boundary for A and H, as well as including a central region of the sigmoid curve for class D. While this did not improve overall accuracy, it did increase the precision for each class which may have some benefit, specifically in a betting context.

Findings and Conclusion

Football is a highly dynamic sport with numerous intertwined factors, including player performance, team dynamics, tactical strategies, and unpredictable chance events on the field. These complexities make predicting football outcomes a challenging task for any model.

Although there have been efforts to improve overall accuracy, the best approach depends on the model's intended use. Some improvements resulted in higher precision for specific classes, but predicting draws remained particularly challenging. The nature of draws as the "middle" value among the classes led to shared similarities with both class A and class H, contributing to the models' limitations.

Despite applying hyperparameter tuning and implementing gradient boosting, the models' overall accuracy only experienced a moderate increase from approximately 46% to 50.5%. Unfortunately, none of the models reached the target accuracy of 55%. However, I believe that achieving this goal is still possible with further feature engineering and potentially incorporating additional data, thereby empowering the models to make more accurate predictions. The complexity of football outcomes necessitates a continual quest for refined models that can better capture the nuances of this unpredictable sport.