Probability Estimation
The Advantage over Bookmakers

Jake Warren
contact.jakewarren@gmail.com

I. Introduction

When it comes to football betting there can be a lot of money to be made, bookmakers each year make hundreds of thousands if not millions on the public betting on football games. With that being said bookmakers use probability to gain advantage over the football games being bet on and this then reflects on the odds that the public get, what if we could gain some advantage back over the bookmakers. With the digital era that we are in there is no shortage of statistics being recorded for these football matches, so being able to gather and leverage this data is becoming easier and if we can estimate the probabilities for the home teams and away teams winning from the data gathered, we can stand a greater chance of gaining an advantage over the bookmakers. It must be said that with any kind of betting there is always an implied risk of losing, oftentimes football teams can go against the run of data that they have, for example, top of the premier league losing or dropping points to bottom of the premier league, it doesn't always happen but it does happen.

Project Outline:
For this project the goal was to estimate the probabilities of winning for the home teams and away teams of football matches in the premier league, championship and league one of english football. The reasoning behind only obtaining probabilities of just winning is because the actual chance of two teams cancelling each other out for a draw is usually unlikely, it is more likely that one team underperforms, or one team overperforms, or both teams are just unlucky on the day. As we cant factor these outcomes in, predicting for a draw would be rather difficult. So the idea for this project was to create two classification models, one for modelling the estimation of home team win probabilities and one for modelling the estimation of away team win probabilities. The highest probability teams for each game are ranked by probability and difference from the team with the lower probability of winning from each game along with their expected value and bet return. These teams then can be bet on and also used in accumulation calculations for accumulator bets. The expected value formulations are as follows;
Home win model:
$P(W \mid ht) * wa + (1 - P(W \mid ht)) * -ba$
Away win model:
$P(W \mid at) * wa + (1 - P(W \mid at)) * -ba$

Where W represents win, ht is the home team, at is the away team, wa represents win amount and ba represents bet amount respectively.

The metric of choice for both models was fbeta with slightly more weight towards maximising recall and minimising false negatives. This was because the distribution of classes for both models were unbalanced more so for the away win model, also because minimising false negatives allows the models to be more risk accepting as decreasing false negatives will imply increasing false positives, in this domain when minimising predicting no win when it actually is a win (false negatives) implies an increase in predicting win when it's actually not a win (false positives). Therefore allowing the model to make more win predictions with the risk of making win predictions when it is actually not win, allowing more risk, although the weighting is only minimal in this project. If fbeta was weighted more towards maximising precision then the models would become more conservative, this is down to choice but also for future analysis a comparison of maximising precision could be used to test if a more conservative model improves probability estimations.

It was found that for a majority class predicting baseline, the home win model produced an F-beta score of approximately 0.42 and the away win model produced an F-beta score of approximately 0.61. A simple model was produced using the most informative feature, both simple home and away win models produced an F-beta score of approximately 0.56. As can be seen due to the bigger class imbalance the away win model F-beta score decreases with the simple model showing more false negative predictions.
With the home win final model we achieved an F-beta score of approximately 0.60 which is an approximate 42% increase over the majority class predicting baseline and an approximate 7% increase over the simple model baseline.
The away win final model achieved an approximate F-beta score of 0.68 which is an approximate 11% increase over the majority class predicting baseline and an approximate 21% increase over the simple model baseline. The top performing final models produced a combined expected profit of £3.74, in the long term on average we can expect £3.74 profit from betting on teams the models estimate betting on.

II. Dataset

The main data were obtained online and is open for the public to use, as the data contains already played match statistics averages had to be computed so we could get predictions for future matches, averages from 5 previous games were chosen. The number of averages could be a parameter to change in future to see if performance can be gained. Some features were engineered using data scraped from websites that allow correlation/association were likely to have occurred by

web scraping and also additional python api's were used to engineer features. The original data collected had 41,382 observations and 41 features after averages were computed, of the 41 features 4 were nominal, 1 was ordinal and 36 were continuous. The target variable full time result 'FTR' was nominal originally with 'H','D','A' representing home win, draw and away win respectively. For the home and away models target engineering was implemented where for the home model a 1 represented home win and 0 represented away win and draw, for the away model 1 represented away win and 0 represented home win and draw. Both engineered targets are unbalanced with the away win target more so, after a train/test split with 20% of data kept for testing the home win proportion was 44% and the away win proportion was 29%. See appendix for full feature list (original and engineered).

III. Preprocessing//Features

Data cleansing initially started with the removal of some features which included 'Referee' as it had a high number of nan values, it also had a high number of unique values so wouldn't provide a lot in the way of information on the targets. Bookmakers odds were removed as well as to gain leverage over the bookmakers by determining from the data which of the home and away teams were likely to win and didn't want these features to influence the models. Although the odds could provide some latent information about the teams and so might be worth including them in future analysis.

A chunk of nan values were removed at the head of the data from the average computed values. The data were also checked for rows containing more than 2 nan values and subsequently were removed, for the rows containing 2 or less nan values similarity imputation was applied using euclidean distance to determine the rows nearest neighbours from which the average value for that feature from the nearest neighbours was applied, also averages were imputed for the team and feature. There were no erroneous or duplicated data and observable outliers could be seen in the form of rare match outcomes but this is a part of football so are kept in, although the majority of the continuous features were right skewed which makes sense as extreme values in football tend to be higher, for example a high number of goals scored is more rare than fewer goals scored. Feature outliers were detected via the interquartile range and addressed separately for numerical function models using Box-Cox, Yeo-Johnson and quantile power transforms to reduce skewness and make more gaussian like, additionally a robust scaler was used to further reduce the influence of outliers for the numerical function models. Also for future reference clipping features to remove outliers whilst providing indication of an outlier could also work. Shapiro, D'agostino and Anderson normality tests were used to identify normal/gaussian distributions with all continuous features showing to be non-gaussian. Spearman's rho, Cramer's v, phi and information gain tests were used to analyse feature/target correlation/association with the least correlated/associated features removed. Chi squared and Spearman's rho statistical tests were also performed for feature/target correlation and association with features being removed if their chance, as they could prove problematic for population generalisation. Initially the data only showed mild multicollinearity but after feature engineering high multicollinearity was introduced and this was addressed separately for numerical function models via the use of PCA to alleviate the presence of collinearity whilst also reduce dimensionality, tree based models are unaffected by multicollinearity so will not be used for these models. Types of features that were engineered include; date extraction (days, months, years), cyclical nature (days, months, years), ratios, differences, conversions, indicators for outliers and less-than or greater-than, binned features, ordinality inversion, power transforms, quartile indicators, feature interactions, quantile indicators, normality transforms and natural forming clusters.

IV. Modelling

As we are modelling the probability estimations of home and away teams winning over not winning we are presented with a binary classification task, and so all models tested were supervised learning algorithms. Before model testing began we acquired a base rate (majority predicting) metric and a simple model (best feature) metric as a comparison for both home and away model learnability. The home model base rate and simple model F-beta scores acquired were 0.4183 and 0.5553 respectively, the away model base rate and simple model F-beta scores acquired were 0.6102 and 0.5644 respectively. It can be seen that the home model performance increases substantially when the top feature is used to predict but the away model performance decreases somewhat. This performance decrease could be due to the bigger class imbalance for the away model with only an approx. 29% away win proportion compared to approx. 44% home win proportion. The decrease could be due to the away simple model predicting more false negatives (predicting away no win when it's actually away win). This can be alleviated by performing oversampling and/or undersampling methods.

As the data were pre processed accordingly for numerical function models we would include these with the other tree based and boosting models being tested, all model testing was performed using 10 fold cross validation.

After the initial model tests a paired T test with 5x2 cross validation was used to check the statistical significance of each models scores and once we had established the home and away models that work well with this data we then proceeded with performing sequential forward feature selection for each model, this would reduce dimensionality for the models whilst also giving a boost to model performance by finding only the optimal features for the problem being solved.As there is an imbalance with the home and away models target class distributions, moreso the away model, undersampling (removing examples of the majority class) and oversampling (creating more examples of the minority class) techniques were used to gain better insights of the feature characteristics for the minority classes, the win class in this case. Without balancing the dataset the models will be better at predicting/estimating the majority class but find it difficult to provide good predictions/estimations of the minority class. The undersampling algorithms implemented

were; [1][2] ***i***. tomek links - remove majority examples at the class borderline that have minority examples as nearest neighbours, ***ii***. edited nearest neighbours - removes misclassified observations, ***iii***. repeated edited nearest neighbours - repeatedly applies edited nearest neighbours until no more observations can be removed, ***iv***. one sided selection - a combination of applying tomek links then applying condensed nearest neighbor method to remove majority examples far away from the class borderline, ***v.*** neighbourhood cleaning rule - combines the condensed nearest neighbor method to remove redundant majority examples and edited nearest neighbor method to remove misclassified majority examples, ***vi***. instance hardness threshold - removes overlapping class observations and ***vii.*** Near miss - NearMiss-1 selects examples from the majority class that have the smallest average distance to the three closest examples from the minority class. NearMiss-2 selects examples from the majority class that have the smallest average distance to the three furthest examples from the minority class. NearMiss-3 involves selecting a given number of majority class examples for each example in the minority class that are closest. The oversampling algorithms implemented were; ***i***. smote - selects minority examples close to each other and creates new examples between them, ***ii***. borderline smote - creates observations from the misclassified observations nearest the class border, ***iii***. svm smote - creates minority observations from misclassified minority observations close to the class decision boundary obtained by the svm algorithm and ***iv***. adaptive synthetic sampling - generates more minority observations in the feature space where a low number of minority observations are found and generates less or no observations where a high number of minority observations are found. A combination of oversampling and undersampling was used using smote enn which combines smote that selects minority examples that are close to each other and creates a new example between the minority examples and edited nearest neighbours to remove misclassified examples from the majority class.

Once an optimal sampling strategy had been chosen for each model we then performed hyperparameter tuning via the means of bayesian optimisation with cross validation. Bayesian optimization uses Bayes Theorem to direct the search in order to find the minimum or maximum of an objective function, in this case F-beta which was turned into a loss score by taking 1 minus the F-beta score, via the use of a surrogate function which estimates the objective function and directs future sampling. Bayesian optimization usually performs more efficiently compared to a grid or random search. Finally as we are using the estimated probabilities of the final model/s adding in model calibration is essential as we want the probabilities to be as close to perfect instead of being over or underconfident. [3] Calibration involves the usage of a calibration regressor in which we have the choice of two;The sigmoid regressor based on platts logistic model:

$$p(yi = 1 \mid fi) = \frac{1}{1 + \exp(Afi + B)}$$

where yi is the true label of sample i and fi is the output of the un-calibrated classifier for sample i. A and B are real numbers to be determined when fitting the regressor via maximum likelihood. The isotonic regressor which fits a non-parametric isotonic regressor which outputs a step-wise non-decreasing function:

$$\sum_{i=1}^{n} \left(yi - \hat{fi} \right)^2$$

Subject to $\hat{fi} \geq \hat{fj}$ whenever $fi \geq fj$. yi is the true label sample i and $\hat{fi}$ is the output of the calibrated classifier for sample i (i.e. the calibrated probability).

V. Results

From initial model testing it was found that the removal of the home and away teams resulted in negligible performance loss whilst gaining computational efficiency, therefore these features were removed. From the data it looked like tree-based and boosting models would show a more accurate performance as the skewness among the features from outliers would not favour numerical function models, but during preprocessing the skewness was reduced allowing numerical function models to be used effectively giving more flexibility with model choice. After Initial model testing *table 1* shows the top performing home models and *table 2* shows the top performing away models.

Table 1 - Top Home Models

Model	F-beta Score
Logistic Regression	0.587389
Gradient Boosting	0.579037
XGBoost	0.578781
Gaussin NB	0.578502
AdaBoost	0.57715
CatBoost	0.577177

Table 2 - Top Away Models

Model	F-beta Score
Light Gradient Boosting	0.645203
XGBoost	0.645075
AdaBoost	0.643416
CatBoost	0.640387
Gradient Boosting	0.639838

A paired T test with 5x2 cross validation was used to determine the significance of the model scores in relation to each model. It was found that the home logistic regression model was statistically better than the rest and all away top models statistically showed the same performance. As can be seen the majority of models are boosting models as assumed would work well but also some numerical function models (logistic regression and gaussian nb) seem to have performed well due to preprocessing for these models. All models out performed the simple and base rate models.
For computational efficiency feature selection was applied to each model using a combination of recursive feature elimination for models that apply feature weighting/importance and select k best for those that don't followed by the sequential forward feature selection method, it was found after, that all models produced approximately the same F-beta score or better, please see appendix for model features and importances. PCA was applied to the numerical function models to reduce dimensionality and multicollinearity. Once the best sampling strategy was found for each model using cross validation followed by hyperparameter tuning, the best models that came out on top were; home models - logistic regression with PCA reduced data to 12 components and near miss version 3 undersampling, XGBoost with borderline smote oversampling and away models - XGBoost with svm smote oversampling, light gradient boosting also with svm smote oversampling.
Figures 1, 2, 3 and 4 show the calibrated probability plots for each model with uncalibrated, sigmoid and isotonic regressors against perfect calibration respectively. As can be seen although the plots show approximately perfect to not far off perfect probabilities the histograms for the plots show that the probabilities being output are not exactly close to 0 or 1. This could be due to the fact that there's a lot of uncertainty when it comes to saying that a team will 100% win or 0% win, for example if the team is the best team in the world then it would be credible to say that they were most likely to win but there are also decisions in the match that could go against them winning i.e. penalties, sendings off, injuries, all these factors would contribute to the best team in the world possibly losing. Therefore it's not unreasonable for the models to show this within their probability predictions and more likely so for teams of average quality.

figure 1-home logistic regression calibration plots

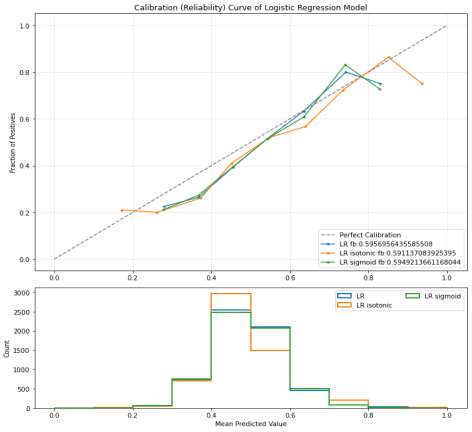


figure 2-home xgboost calibration plots

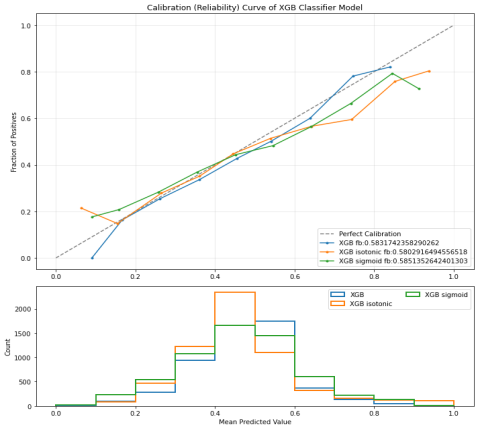


figure 3-away light gradient boosting calibration plot

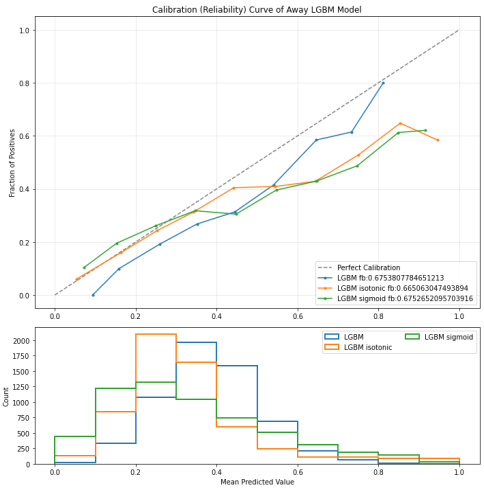
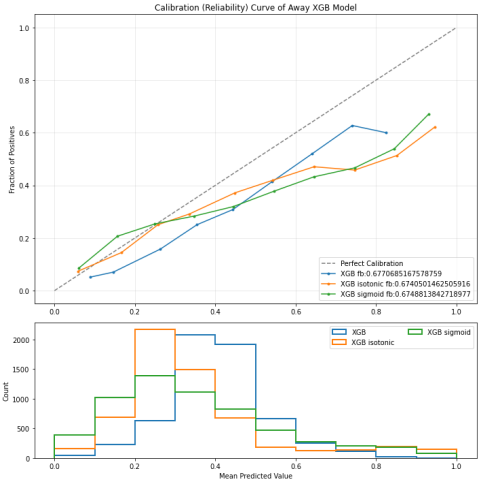


figure 4-away xgboost calibration plot



Along with the individual models a voting classifier, equal weighted ensemble and optimised weighted ensemble will be put forward for final testing to see if any performance can be

gained from combining the individual models.
Figures 5 through 10 show model learning curves, roc curves and cumulative response curves for home, away, voting classifier and ensemble models respectively. The learning curves for the home models show that by obtaining more data we are likely to see an improvement on performance, the logistic regression model appears to have more variability with its cross validated scores. The learning curve for the away XGBoost model shows a plateau leading to a slight decline with the test score meaning that more data might not possibly improve the F-beta score. The learning curve for the away light gradient boosting model shows the test score starting to plateau which also could mean more data might not improve the F-beta score. The reason behind the plateau and decline could be that the proportion of away wins is fairly low at approx. 29% and there's not many good predictors of an away win, which leads to the increase of data not having an effect on away win prediction performance. Also because an away win is harder to predict we will run into more false positive and false negative predictions, this could be because the home/away team either over perform or underperform, dip in player confidence and other underlying factors. To boost away win prediction performance and improve the learning curves more attention to feature engineering for this area will help.

figure 5-home win model learning curves

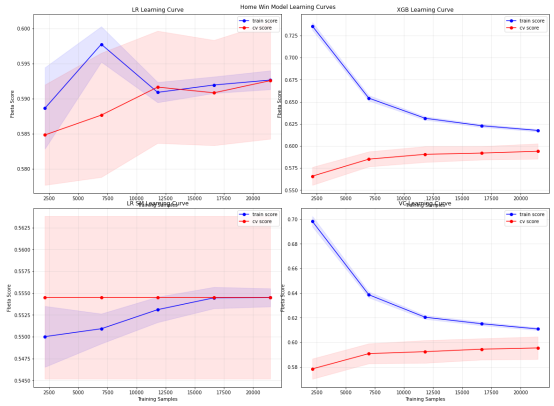
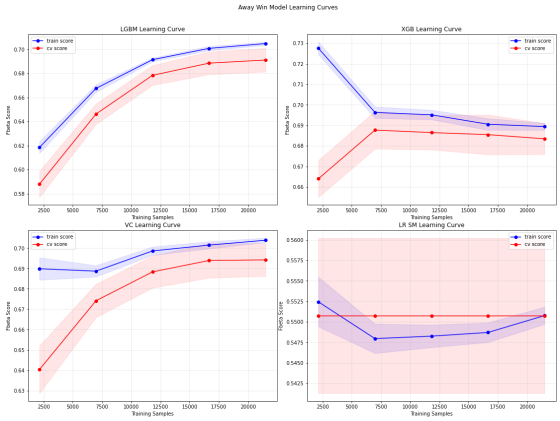


figure 6-away win model learning curves



All home models tested show ROC curves better than random and better compared to the simple model, with AUC scores of approx 0.62. The top AUC scores are shown by the equal weighted ensemble and the optimised weighted ensemble. The home model ROC curves are all approximately the same with the Logistic regression and optimal weighted ensemble having a slight peak at approx. 55% true positive rate to 35% false positive rate. All away models tested show ROC curves better than random and better than the simple model with curves all approx. the same. The optimised weighted ensemble has the best AUC score at 0.6439 just above the equal weighted ensemble.

figure 7-home win model ROC curves

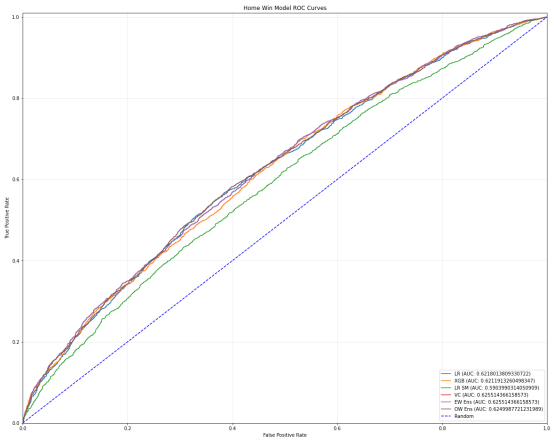
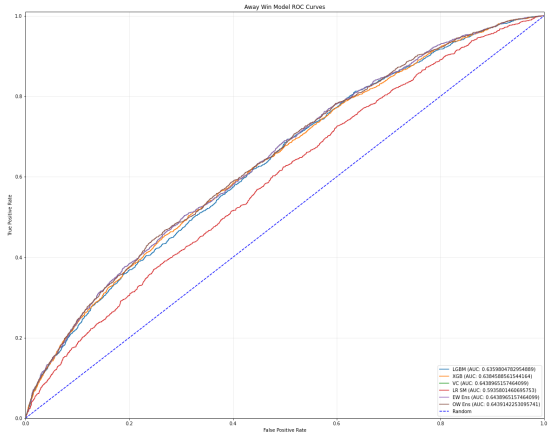


figure 8-away win model ROC curves



The home model cumulative response curves show that the logistic regression, equal weighted ensemble and optimised weighted ensemble produce roughly the same true positive rate for the percentage of data seen. The equal weighted ensemble does have a slight increase at around 60 percent of the data albeit very minimal. All away model cumulative response curves are roughly the same up until approx. 40% of data seen then the simple models true positive rate exceeds the other models. For both home and away models the optimized weighted ensemble proved to pull ahead the rest of

the models, interestingly the simple models were added into the optimised ensembles and increased performance again. This could possibly mean that including some of the other models will increase performance further still and should be taken into consideration in the future.

figure 9-home win model cumulative response curves

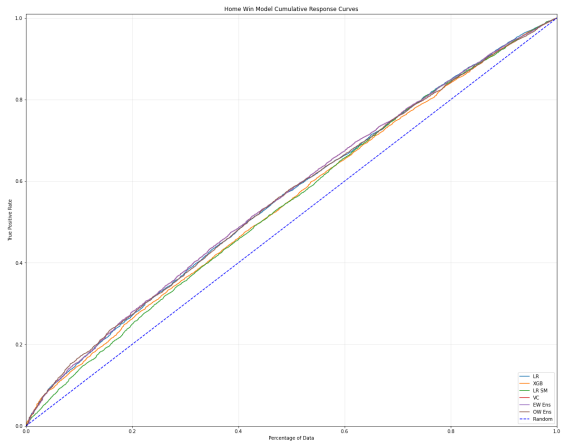
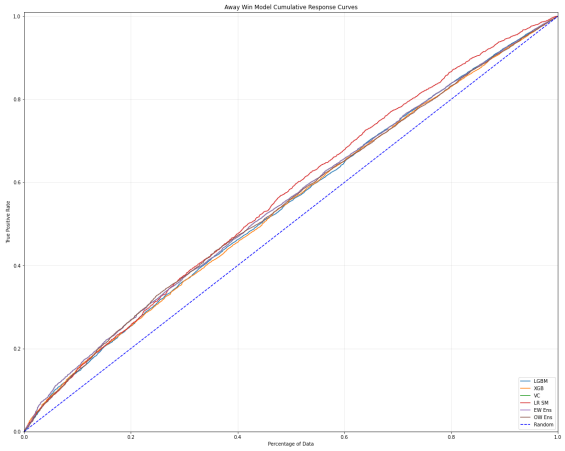


figure 10-away win model cumulative response curves



Finally the profitability of the models was taken into consideration in producing expected profits for the combined home and away models, as they are going to be used in conjunction with each other. Calculated using an approximated average return from a £5 bet (expected profit is likely to change if the actual odds are used in the calculation rather than an average but using an average does allow for comparison), the top expected profit of £3.93 was produced by the top home ensemble model and the away simple logistic regression model. In comparison to the two top metric performing home and away ensembles producing an expected profit of £3.74. It is clear why the away simple model gives a boost to expected profit from the away cumulative response curve showing higher true positive rate from 40 percent of data onwards, but looking at the confusion matrix of the away simple model it also shows to be extremely risk accepting showing more than three times the number of false positives and just over half of the false negatives compared to the other away models. This does show that allowing more risk can be beneficial with more rewards but with this many false positives also can be untrustworthy which is backed up with a cohen kappa score of half that of the top away ensemble model. With this in mind the best expected profit was that of the top home and away ensemble models. Looking at the confusion matrices of the home models and knowing that minimising false negatives leads to a bigger return we can afford to apply more weight towards recall as the false negatives are not far off of the false positives, this will enable the home models to produce a better expected profit. The away model confusion matrices except the simple model show more false negatives than false positives so minimising recall should produce better expected profit, allow more risk and provide better estimations.

VI. Conclusion

The goal of this project was to leverage data via the use of machine learning and data analysis to gain an advantage over the bookmakers with an increase on the number of returns from bets. Probability estimation was the main focus to provide as close to accurate probabilities for home and away team win percentages, these probabilities will be used in calculating the expected values of the teams to be chosen for betting and the expected profit of the models. From the outset it was already understood that predicting football outcomes is reliant on many factors and just because you are the best team is no guarantee that you'll win, with this in mind feature engineering would play an important role in trying to find features that could provide good insights to predicting the targets. We found through the acquisition of football ground capacities that the difference between the away team and home teams stadium capacities proved to be a strong feature, this would also be due to latent information that comes with stadium size such as more money, better facilities and possibly better players. We also found 4 unidentified clusters within the data and two of the clusters also proved to be valuable for the models. We started off with majority predicting baseline models as well as a simple model with the most informative feature 'AwayCapacityDiff' for home and away win models, home win models achieved a majority F-beta of 0.4183 and simple F-beta off 0.5553, away win models achieved a majority F-beta of 0.6102 and simple F-beta of 0.5644. Compared to the final models which were both optimised weighted ensembles of the two best home and away models including the simple models, the home win final model achieved an F-beta score of approximately 0.60 which is an approximate 42% increase over the majority class predicting baseline and an approximate 7% increase over the simple model baseline. The away win final model achieved an approximate F-beta score of 0.68 which is an approximate 11% increase over the majority class predicting baseline and an approximate 21% increase over the simple model baseline. The two optimised weighted home and away ensembles produced an expected profit of £3.74, so in the long term on average we can expect this profit on every bet we take from the models predictions.

Both the top models produced a cohen kappa score of approximately 0.18 which was the best among the models tested but is not a trustworthy score, this along with the expected profit reflects the variability among the outcomes of football matches and with gambling in general.
The proposed usage of the models is to estimate the probabilities of home and away wins for each game, take the difference $(\max(prob) - \min(prob))$, the bigger the difference should indicate more chance of winning and also take the team with the highest probability, calculate expected value and order the teams with respect to probability, difference in probability and expected value in order to choose teams to bet on. Additionally the probabilities will be used to calculate accumulation bets in the same manner as individual bets. In comparison this should prove to be more effective than just looking at 5 previous game win, lose or draw form and league positions. The proof will be in the number of returns from bets.

VII. Future Improvements

As indicated at the beginning the odds of bookmakers were removed with the idea as to not influence the models and gain no advantage. However the odds provided might hold key latent information that is not included in the data such as key players missing the match, this may prove to be beneficial as it will not be found within the data itself. Also these odds can be used in calculating expected profits, which in turn could provide beneficially a better analysis of model performance and profit return compared to using averaged bet returns.
If it is possible to identify the four natural forming clusters we could find better insights to help determine whether the home or away team is more likely to win, also it could lead to more informative feature engineering. Over the course of the global pandemic it has been seen that without fans in stadiums there has been a somewhat leveling of playing fields between lower quality teams and higher quality teams with more ‘upsets’ in football matches, research into fans interactions with their teams could provide essential insights. This also raises the question of whether the last season and a half with no fans will affect future modelling as it seems to affect the conditions and outcomes of football matches. As outlined the weighting of the F-beta metric is weighted as such to be more risk accepting, this is due to choice and it was found in model analysis that more risk does in turn produce more return but there is a fine line. Looking at home models more weighting towards recall will be beneficial and away models should directly maximise recall for more benefits in making more correct estimations.
The inclusion of more models in the final ensemble models could prove to also aid performance.

References

[1]. *www.imbalanced-learn.org*
[2].*www.machinelearningmastery.com/data-sampling-methods-for-imbalanced-classification/*
[3]. *https://scikit-learn.org/stable/modules/calibration.html*

Appendix

Full feature list (original and engineered):

Feature	**Description**	**Type**	**Original Feature**
'Div'	Division - (premier, championship, league 1)	Nominal converted to ordinal	yes
'HomeTeam'	name of home team	Nominal	yes
'AwayTeam'	name of away team	Nominal	yes
'FTHG'	Full time home goals	Discrete continuous	yes
'AHTGS5PG'	Average home team goals scored 5 previous games	continuous	no
'FTAG'	Full time away goals	Discrete continuous	yes
'AATGS5PG'	Average away team goals scored 5 previous games	continuous	no
'AHTGC5PG'	Average home team goals conceded 5 previous games	continuous	no
'AATGC5PG'	Average away team goals conceded 5 previous games	continuous	no
'AHTGS5PHG'	Average home team goals scored 5 previous home games	continuous	no
'AATGS5PAG'	Average away team goals scored 5 previous away games	continuous	no
'AHTGC5PHG'	Average home team goals conceded 5 previous home games	continuous	no
'AATGC5PAG',	Average away team goals conceded 5 previous away games	continuous	no
'HS'	Home shots	discrete	yes
'AS'	Away shots	discrete	yes
'HST'	Home shots on target	discrete	yes
'AST'	Away shots on target	discrete	yes
'AHTSOT5PG'	Average home team shots on target 5 previous games	continuous	no
'AATSOT5PG'	Average away team shots on target 5 previous games	continuous	no

Feature	Description	Type	Original Feature
'AHTSOT5PHG'	Average home team shots on target 5 previous home games	continuous	no
'AATSOT5PAG'	Average away team shots on target 5 previous away games	continuous	no
'HF'	Home team freekicks	discrete	yes
'AF'	Away team freekicks	discrete	yes
'HC'	Home team corners	discrete	yes
'AC'	Away team corners	discrete	yes
'HY'	Home team yellow cards	discrete	yes
'AY'	Away team yellow cards	discrete	yes
'HR'	Home team red cards	discrete	yes
'AR',	Away team red cards	discrete	yes
'AHTP5PG'	Average home team points 5 previous games	continuous	no
'AATP5PG'	Average away team points 5 previous games	continuous	no
'AHTP5PHG'	Average home team points 5 previous home games	continuous	no
'AATP5PAG',	Average away team points 5 previous away games	continuous	no
'month',	Month of year (integer)	discrete	no
'year',	year	discrete	no
'DayofWeek',	Day of week (integer)	discrete	no
'AHTGS_SOT5PG_ratio',	Average home team goals scored 5 previous games divided by average home team shots on target 5 previous games	continuous	no
'AATGS_SOT5PG_ratio',	Average away team goals scored 5 previous games divided by average away team shots on target 5 previous games	continuous	no
'AHTGS_SOT5PHG_ratio'	Average home team goals scored 5 previous home games divided by average home team shots on target 5 previous home games	continuous	no
'AATGS_SOT5PAG_ratio'	Average away team goals scored 5 previous away games divided by average away team shots on target 5 previous away games	continuous	no
'AHTGD5PG'	Average home team goal difference 5 previous games - (AHTGS5PG - AHTGC5PG)	continuous	no

Feature	Description	Type	Original Feature
'AATGD5PG'	Average away team goal difference 5 previous games - (AATGS5PG - AATGC5PG)	continuous	no
'AHTGD5PHG'	Average home team goal difference 5 previous home games - (AHTGS5PHG - AHTGC5PHG)	continuous	no
'AATGD5PAG'	Average away team goal difference 5 previous home games - (AATGS5PHG - AATGC5PHG)	continuous	no
'season_month'	Month of season (beginning month = 1)	discrete	no
'season_month_sin'	Sine transformation for cyclical nature	cyclical	no
'season_month_cos'	Cosine transformation for cyclical nature	cyclical	no
'DayofWeek_sin'	Sine transformation for cyclical nature	cyclical	no
'DayofWeek_cos'	Cosine transformation for cyclical nature	cyclical	no
'AHTGS5PG_UPoutlier'	Indicator for upper outliers	dichotomous	no
'AATGS5PG_UPoutlier'	Indicator for upper outliers	dichotomous	no
'AATGS5PG_LOWoutlier'	Indicator for lower outliers	dichotomous	no
'AHTGC5PG_UPoutlier'	Indicator for upper outliers	dichotomous	no
'AHTGC5PG_LOWoutlier'	Indicator for lower outliers	dichotomous	no
'AATGC5PG_UPoutlier'	Indicator for upper outliers	dichotomous	no
'AHTGS5PHG_UPoutlier'	Indicator for upper outliers	dichotomous	no
'AATGS5PAG_UPoutlier'	Indicator for upper outliers	dichotomous	no
'AHTGC5PHG_UPoutlier'	Indicator for upper outliers	dichotomous	no
'AATGC5PAG_UPoutlier'	Indicator for upper outliers	dichotomous	no
'AHTSOT5PG_UPoutlier'	Indicator for upper outliers	dichotomous	no
'AHTSOT5PG_LOWoutlier'	Indicator for lower outliers	dichotomous	no
'AATSOT5PG_UPoutlier'	Indicator for upper outliers	dichotomous	no
'AATSOT5PG_LOWoutlier'	Indicator for lower outliers	dichotomous	no
'AHTSOT5PHG_UPoutlier'	Indicator for upper outliers	dichotomous	no
'AATSOT5PAG_UPoutlier'	Indicator for upper outliers	dichotomous	no
'AATSOT5PAG_LOWoutlier'	Indicator for lower outliers	dichotomous	no

Feature	Description	Type	Original Feature
'AHTGS_SOT5PG_ratio_UP outlier'	Average home team goals scored to shots on target 5 previous games ratio (AHTGS5PG / AHTSOT5PG)	continuous	no
'AATGS_SOT5PG_ratio_UP outlier'	Average away team goals scored to shots on target 5 previous games ratio (AATGS5PG / AATSOT5PG)	continuous	no
'AHTGS_SOT5PHG_ratio_ UPoutlier'	Average home team goals scored to shots on target 5 previous home games ratio (AHTGS5PHG / AHTSOT5PHG)	continuous	no
'AATGS_SOT5PAG_ratio_U Poutlier'	Average away team goals scored to shots on target 5 previous away games ratio (AATGS5PAG / AATSOT5PAG)	continuous	no
'AHTGD5PG_UPoutlier'	Average home teams goal difference 5 previous games upper outlier indicator	dichotomous	no
'AHTGD5PG_LOWoutlier'	Average home teams goal difference 5 previous games lower outlier indicator	dichotomous	no
'AATGD5PG_UPoutlier',	Average away teams goal difference 5 previous games upper outlier indicator	dichotomous	no
'AATGD5PG_LOWoutlier'	Average away teams goal difference 5 previous games lower outlier indicator	dichotomous	no
'AHTGD5PHG_UPoutlier'	Average home teams goal difference 5 previous home games upper outlier indicator	dichotomous	no
'AHTGD5PHG_LOWoutlier'	Average home teams goal difference 5 previous home games lower outlier indicator	dichotomous	no
'AATGD5PAG_UPoutlier'	Average away teams goal difference 5 previous away games upper outlier indicator	dichotomous	no
'AATGD5PAG_LOWoutlier'	Average away teams goal difference 5 previous away games lower outlier indicator	dichotomous	no
'HA_AHTGS5PG_diff'	Home and away team average goals scored 5 previous games difference (AHTGS5PG - AATGS5PG)	continuous	no
'HA_ATP5PG_diff'	Home and away team average points 5 previous games difference (AHTP5PG - AATP5PG)	continuos	no
'AHT_GS_P5PG_ratio'	Home team average goals scored to average points ratio (AHTGS5PG / AHTP5PG)	continuous	no
'AAT_GS_P5PG_ratio'	Away team average goals scored to average points ratio (AATGS5PG / AATP5PG)	continuous	no
'AwayTeamDist'	Away team distance from home team (miles)	continuous	no
'AwayCapacityDiff'	Away team capacity difference from home team	discrete	no

Feature	Description	Type	Original Feature
'AwayCapacityDiff_bin'	Away team capacity difference from home team binned	discrete	no
'AwayTeamDist_bin'	Away team distance from home team (miles) binned	discrete	no
'Local_Derby',	Indicator for distance below or equal to 10 miles	dichotomous	no
'Dist>=100'	Indicator for distance above or equal to 100 miles	dichotomous	no
'cluster_0'	Naturally formed cluster	dichotomous	no
'cluster_1'	Naturally formed cluster	dichotomous	no
'cluster_2'	Naturally formed cluster	dichotomous	no
'cluster_3'	Naturally formed cluster	dichotomous	no
'HT_PrevSeasonPos_inv'	Home team previous season finishing position inverted for ordinality	ordinal	no
'AT_PrevSeasonPos_inv'	Away team previous season finishing position inverted for ordinality	ordinal	no
'bxcx_AATGS5PG'	Box-Cox power transformation	continuous	no
'bxcx_AHTGS5PG'	Box-Cox power transformation	continuous	no
'bxcx_AHTGC5PG'	Box-Cox power transformation	continuous	no
'bxcx_AHTSOT5PG'	Box-Cox power transformation	continuous	no
'bxcx_AATSOT5PG'	Box-Cox power transformation	continuous	no
'bxcx_AHT_GS_P5PG_ratio'	Box-Cox power transformation	continuous	no
'bxcx_AAT_GS_P5PG_ratio'	Box-Cox power transformation	continuous	no
'bxcx_AATGC5PG'	Box-Cox power transformation	continuous	no
'HA_AHTGS5PG_diff_upqrt'	Home and away team goals scored 5 previous games difference upper quartile indicator	dichotomous	no
'HA_AHTGS5PG_diff_lowqrt'	Home and away team goals scored 5 previous games difference lower quartile indicator	dichotomous	no
'AHTGC5PG_upqrt'	Average home team goals conceded 5 previous games upper quartile indicator	dichotomous	no
'AHTGC5PG_lowqrt'	Average home team goals conceded 5 previous games lower quartile indicator	dichotomous	no
'AATGC5PG_upqrt'	Average away team goals conceded 5 previous games upper quartile indicator	dichotomous	no

Feature	Description	Type	Original Feature
'AATGC5PG_lowqrt'	Average away team goals conceded 5 previous games lower quartile indicator	dichotomous	no
'HT_GSGC_UPLOW_QRT'	Indicator for AHTGS5PG upper quartile and AHTGC5PG lower quartile	dichotomous	no
'AT_GSGC_UPLOW_QRT',	Indicator for AATGS5PG upper quartile and AATGC5PG lower quartile	dichotomous	no
'AHTGS5PG_upqrt_AATGC5PG_lowqrt',	Indicator for AHTGS5PG upper quartile and AATGC5PG lower quartile	dichotomous	no
'AATGS5PG_upqrt_AHTGC5PG_lowqrt',	Indicator AATGS5PG upper quartile and AHTGC5PG lower quartile	dichotomous	no
'AHTGS5PG_upqrt_AATGS5PG_lowqrt'	Indicator for AHTGS5PG upper quartile and AATGS5PG lower quartile	dichotomous	no
'AATGS5PG_upqrt_AHTGS5PG_lowqrt',	Indicator for AATGS5PG upper quartile and AHTGS5PG lower quartile	dichotomous	no
'AHTGS5PHG_upqrt_AATGS5PAG_lowqrt'	Indicator for AHTGS5PHG upper quartile and AATGS5PAG lower quartile	dichotomous	no
'AATGS5PAG_upqrt_AHTGS5PHG_lowqrt'	Indicator for AATGS5PAG upper quartile and AHTGS5PHG lower quartile	dichotomous	no
'AHTSOT5PG_upqrt_AATSOT5PG_lowqrt'	Indicator for AHTSOT5PG upper quartile and AATSOT5PG lower quartile	dichotomous	no
'AATSOT5PG_upqrt_AHTSOT5PG_lowqrt'	Indicator for AATSOT5PG upper quartile and AHTSOT5PG lower quartile	dichotomous	no
'AHTSOT5PHG_upqrt_AATSOT5PAG_lowqrt'	Indicator for AHTSOT5PHG upper quartile and AATSOT5PAG lower quartile	dichotomous	no
'AATSOT5PAG_upqrt_AHTSOT5PHG_lowqrt'	Indicator for AATSOT5PAG upper quartile and AHTSOT5PHG lower quartile	dichotomous	no
'AHTGC5PG_upqrt_AATGC5PG_lowqrt'	Indicator for AHTGC5PG upper quartile and AATGC5PG lower quartile	dichotomous	no
'AATGC5PG_upqrt_AHTGC5PG_lowqrt'	Indicator for AATGC5PG upper quartile and AHTGC5PG lower quartile	dichotomous	no
'HTbigcapacitydiff_highgs5pg_lowgc5pg'	Indicator for AwayCapacityDiff_bin below 0.2 quantiles, AHTGS5PG >= 2.2 & AHTGC5PG < 0.8	dichotomous	no

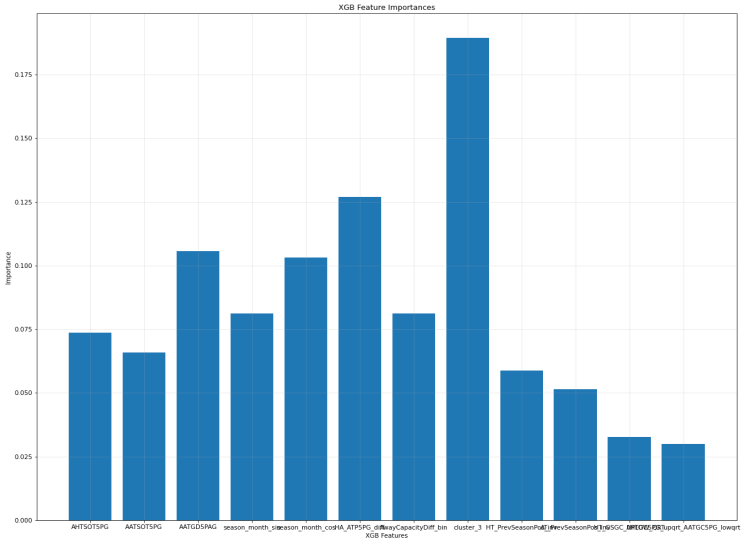
Continuous transformations for numerical function models:

'AHTGS5PG_quantileTRANSFORM', 'AATGS5PG_quantileTRANSFORM', 'AHTGC5PG_quantileTRANSFORM', 'AATGC5PG_quantileTRANSFORM', 'AHTGS5PHG_quantileTRANSFORM', 'AATGS5PAG_quantileTRANSFORM', 'AHTGC5PHG_quantileTRANSFORM', 'AATGC5PAG_quantileTRANSFORM', 'AHTSOT5PG_bxcx_pwrTRANSFORM', 'AATSOT5PG_bxcx_pwrTRANSFORM', 'AHTSOT5PHG_bxcx_pwrTRANSFORM', 'AATSOT5PAG_bxcx_pwrTRANSFORM', 'AHTP5PG_bxcx_pwrTRANSFORM', 'AHTGS_SOT5PG_ratio_quantileTRANSFORM', 'AATGS_SOT5PG_ratio_quantileTRANSFORM', 'AHTGS_SOT5PHG_ratio_bxcx_pwrTRANSFORM', 'AATGS_SOT5PAG_ratio_quantileTRANSFORM', 'AHTGD5PG_quantileTRANSFORM', 'AATGD5PG_quantileTRANSFORM', 'AHTGD5PHG_quantileTRANSFORM', 'AATGD5PAG_quantileTRANSFORM', 'HA_AHTGS5PG_diff_quantileTRANSFORM', 'HA_ATP5PG_diff_quantileTRANSFORM', 'AHT_GS_P5PG_ratio_quantileTRANSFORM', 'AAT_GS_P5PG_ratio_quantileTRANSFORM', 'AwayTeamDist_quantileTRANSFORM', 'AwayCapacityDiff_bin_quantileTRANSFORM', 'AwayTeamDist_bin_quantileTRANSFORM', 'bxcx_AATGS5PG_quantileTRANSFORM', 'bxcx_AHTGS5PG_quantileTRANSFORM', 'bxcx_AHTGC5PG_quantileTRANSFORM', 'bxcx_AHTSOT5PG_quantileTRANSFORM', 'bxcx_AATSOT5PG_quantileTRANSFORM', 'bxcx_AHT_GS_P5PG_ratio_quantileTRANSFORM', 'bxcx_AAT_GS_P5PG_ratio_quantileTRANSFORM', 'bxcx_AATGC5PG_quantileTRANSFORM'

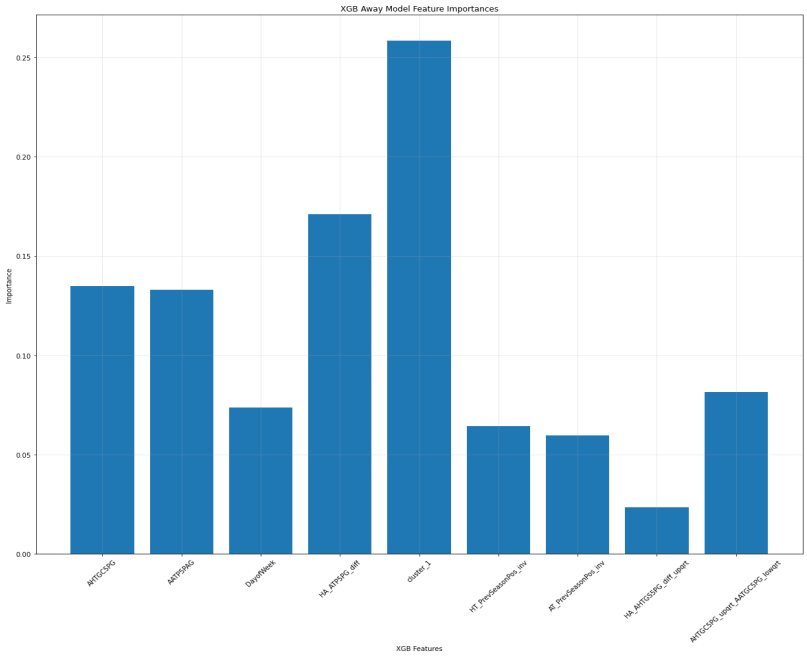
Model Features and Importances

Home logistic regression model features:
'AHTGS5PG_UPoutlier', 'AHTGS5PHG_UPoutlier', 'AATGS5PAG_UPoutlier', 'AHTSOT5PG_UPoutlier', 'AATSOT5PG_UPoutlier', 'AHTSOT5PHG_UPoutlier', 'AHTGD5PHG_UPoutlier', 'HT_PrevSeasonPos_inv', 'AT_PrevSeasonPos_inv', 'HA_AHTGS5PG_diff_lowqrt', 'AHTGS5PG_upqrt_AATGC5PG_lowqrt', 'AHTSOT5PG_bxcx_pwrTRANSFORM', 'AATSOT5PG_bxcx_pwrTRANSFORM', 'AHTGD5PG_quantileTRANSFORM', 'AATGD5PG_quantileTRANSFORM', 'AwayCapacityDiff_bin_quantileTRANSFORM', 'bxcx_AHTSOT5PG_quantileTRANSFORM', 'bxcx_AATSOT5PG_quantileTRANSFORM', 'bxcx_AAT_GS_P5PG_ratio_quantileTRANSFORM'

Home xgboost model features and importances:
'AHTGC5PHG', 'AATSOT5PG', 'AHTSOT5PHG', 'HA_AHTGS5PG_diff', 'AHT_GS_P5PG_ratio', 'AwayCapacityDiff_bin', 'cluster_1', 'cluster_3', 'HT_PrevSeasonPos_inv', 'AT_PrevSeasonPos_inv', 'bxcx_AHTGC5PG', 'bxcx_AATSOT5PG', 'bxcx_AHT_GS_P5PG_ratio'



Away xgboost model features and importances:
'AHTGC5PG', 'AATP5PAG', 'DayofWeek', 'HA_ATP5PG_diff', 'cluster_1', 'HT_PrevSeasonPos_inv',
'AT_PrevSeasonPos_inv', 'HA_AHTGS5PG_diff_upqrt', 'AHTGC5PG_upqrt_AATGC5PG_lowqrt'



Away light gradient boosting model features and importances:
'AHTGS5PG', 'AHTGC5PG', 'AHTSOT5PHG', 'AATSOT5PAG', 'AHTP5PHG', 'AHTGD5PG',
'HA_AHTGS5PG_diff', 'HA_ATP5PG_diff', 'AHT_GS_P5PG_ratio', 'AwayCapacityDiff_bin',
'HT_PrevSeasonPos_inv', 'AT_PrevSeasonPos_inv'