Predicting Corners

Balaram Sridhar

Version updates

V1: Modelling and validation documentation

V2. Sections on applying to betting lines (end of Page 6, Page 17,18)

Overview of problem and summary

The objective is to use the training data provided to predict the number of corners for the matches in the test set, and then use these predictions along with the odds and lines provided to identify which matches to bet on and how much to bet. This problem can be split into the following parts

1. Data - preparing and analysing the training and test data to prep the data to train the model(s) and decide what kind of modelling methodologies may be effective.
2. Feature Selection - checking the features available and their relation to the response variables, any feature engineering, checking for any interactions to fit (in the case of GLM models).
3. Model building and validation - fitting the models, validating the model performance using A vs E charts, Double lift charts and other validation metrics.
4. Bet allocation - deriving implied probabilities, identifying where expected payout is positive and optimising the amount to bet.

Quick model summary

Response variable: Total Corners = Home corners + Away Corners. After preliminary analysis, Negative binomial distribution chosen due to overdispersion and a support including 0.

Model: GLM predicting total number of corners trained on 80% of the training data provided, with the remaining 20% used for validation.

3 models were fitted - one with all the features, one with features excluding features that showed a flat trend in the one-way charts, and lastly one with interactions.

Backward selection was used, using standard errors and AICc to evaluate the benefit of adding/removing factors. Factors were also removed based on one-way charts

Features: Final features used in the model after pruning and fitting interactions are in the attached pdf with brief justifications.

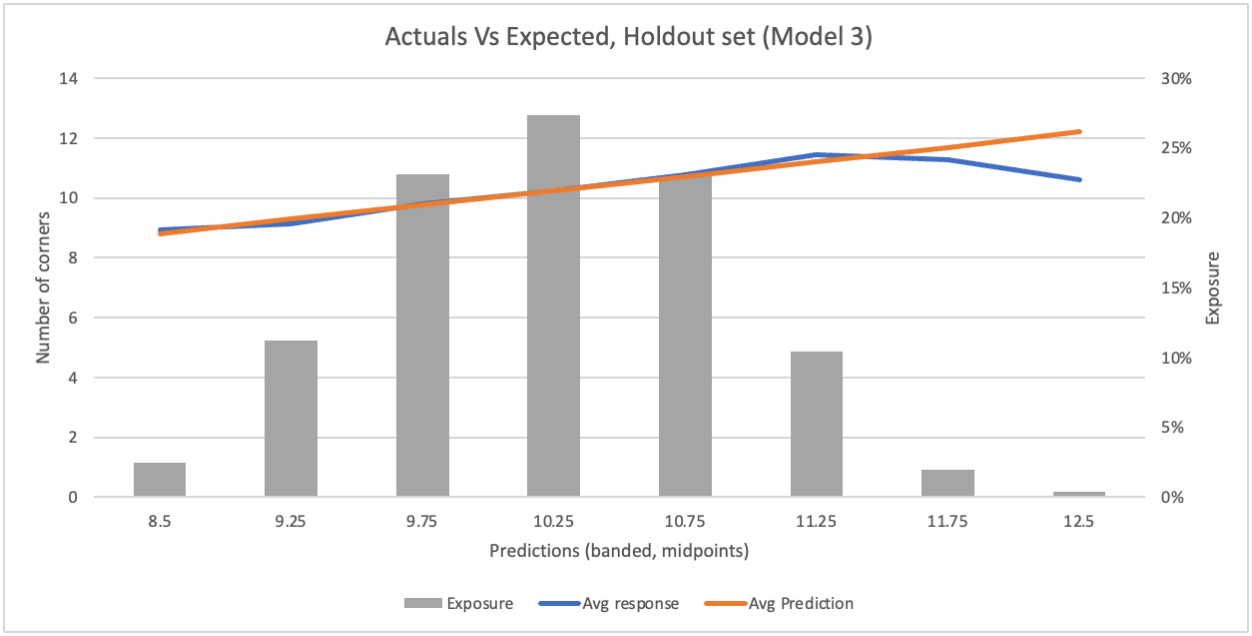
In general there are 3 categories: Goals and corners, Team form and League features. These are mostly rolling averages joined on by match ID and either team or league ID. For the test set, I joined on the latest of these values for each team and league. The rolling average periods were 5,10 and 50 respectively, since typically one would expect strategy/playing style to change more frequently than team form (hence 5 vs 10), and league features can allow for more matches to get baseline averages.

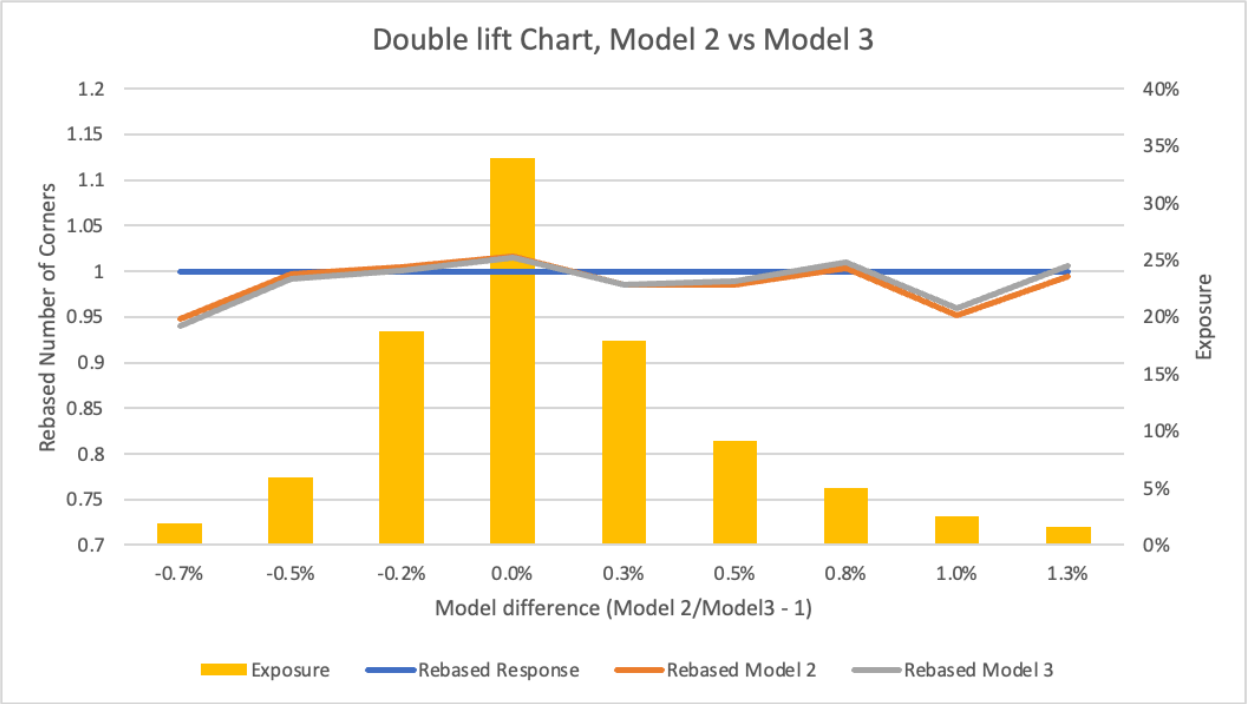
Interactions were fitted, and double lift chart in model validation showed that they add value to the model performance.

Validation results: Actuals vs Expected charts showed that all three models generalise well, with a good fit on the holdout set. Some gapping was observed at there upper end, however this most likely volatility due to an extremely low volume of data. This gapping was larger in model 1 than 2, and larger in 2 than 3 and this was also validated using RMSE (lowest for 3, then 2 then 1), and an exposure weighted distance metric: (Avg prediction - Avg response)^2*exposure then summed over all the bands.

Double lift charts showed that on the holdout set, Model 3 tracked better to the observed than Model 2 (grey closer to blue line in 2nd chart below).

Hence Model 3 was chosen as the best model of the 3 and used to predict on the test data provided.





Betting selection and stake: Using the the dispersion parameter found in preliminary analysis, and the predictions from Model 3 as the mean, we can find the parameters for a Negative Binomial distribution work out the implied probabilities of being under/at/over the given line (all calculations/formulas in the completed excel file). For non-push lines, P(at) = 0, and then the expected profit (EP) of betting over and EP of betting under can be calculated. If both negative no bet was placed, otherwise the one giving highest EP would be selected.

Betting stake can be optimised also by working out EP after one bet where total wealth is 1, then working out the best proportion of 1 to bet (calculation in final section of this document). Then rescale to ensure the sum bet is equal to the total amount available to stake.

Other considerations

Modelling - Next steps would include a few other options could be tested and compared, including GLMs with elastic net regularisation, a GLM + GBM boost structure, GBM only. Another possible approach could be to add a very small positive number (eg. 0.000000001) to every row for each of the 3 variables and then check if a Gamma distribution can be used, since this requires a positive support. I have chosen to treat the variables as NB distributed.

Features - different periods of rolling averages can be tested, eg ratio of avg corners over last 5 to avg corners in last 10 matches. With more time I'd also like to investigate more interactions, eg

Avg corners in last 5 with Avg corner difference since corner difference ignores number of corners and you could have very end-to-end games not being picked up without the interaction. Lastly, although factors were removed if they showed little trend with the response variable, I would like to further review the factor selection since some factors may be linked and the model may benefit from removing some more - eg. league average home/away goals in last 50 and league avg total goals in last 50. Although the model fitting summary indicates low standard errors and all 3 show strong trends in one-way charts, it could be worth a more detailed investigation.

Average Total corners in last N matches for home and away teams is a factor that I realised might be useful too late on. This indicates if the team is involved in higher corner matches, and using different periods for the average could indicate how consistent this is.

Details

Modelling Methodology

There are two main considerations in deciding the modelling methodology.

1. Modelling total corners - a single model can be built to predict the total number of corners immediately, and use this to work out the implied probability of being under/at/over.
2. Modelling Home Corners and Away corners separately - This may be beneficial since home and away behaviours are different, so modelling them separately captures this difference in a more granular manner. It is also easier to identify which features drive home vs. away corner generation, however this method ignores any correlation between home and away corners.

For speed I have chosen to build GLMs for both cases, which I can then compare and choose the better performing structure using double lift charts.

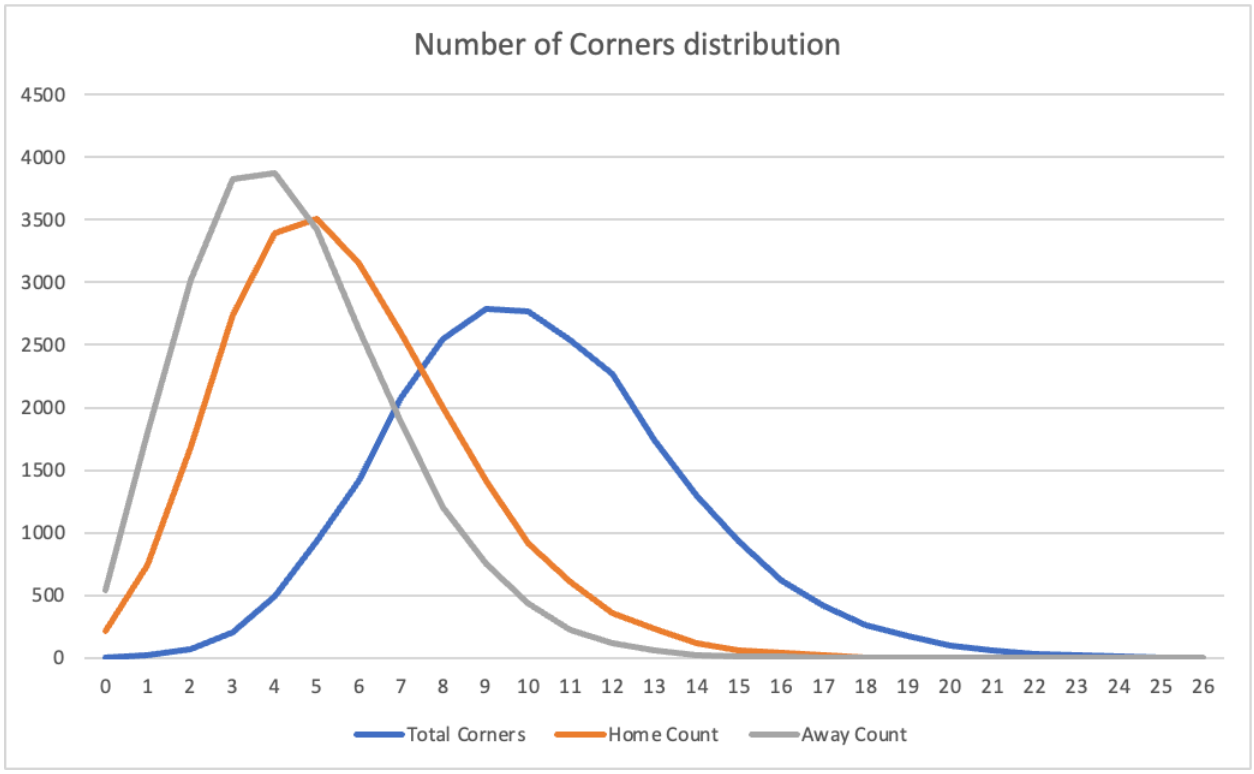
Data

Response Variable

Corners are a count variable which typically follow a Poisson distribution, however when I checked the distribution of Total, Home and Away corners in the training set I saw that they all showed some positive skew, and in all cases variance was higher than the mean. This can be seen in the below image.

	Mean	Variance	r	p
Total	10.21947125	11.9159917	61.55988	0.85762659
Home	5.726059589	8.09239653	13.8559128	0.70758515
Away	4.493411666	6.27690904	11.3208736	0.71586375

This suggest that a Poisson distribution might be too tight at the tails since Poisson distribution requires that the mean and variance are broadly equal, and visually from the below chart the 3 variables look to be closer to a Negative Binomial (NB) distribution. A Chi-Square goodness of fit test was performed using the worked out parameters for the NB distribution (r and p), showing that the 3 variables can be modelled using an NB distribution.



Features

The features given in the training dataset are

Feature	Data type	Use
League ID	Int	Different leagues may show different behaviours, eg. could be useful to compare teams corners to league average
Date	Date	Can be used to generate rolling averages
Home Team ID	Int	To create lookup factors
Away Team ID	Int	To create lookup factors
Home Team Goals	Int	Intuitively, higher scoring team and higher conceding teams a likely to be involved in matches with a higher number of corners.
Away Team Goals	Int	Intuitively, higher scoring team and higher conceding teams a likely to be involved in matches with a higher number of corners.

Home team and Away team corners were also provided which intuitively are linked to Total corners, and these will also be used to create features that can be looked up using the League and Team IDs. The only features available at the point of prediction (in the test dataset) are the league ID, Team IDs and the data. This means that lookup factors will need to be generated to incorporate the goals and corners features, eg.average home team goals scored, rolling averages over last N matches.

Generated factors from this list will be covered in the feature engineering section.

Cleaning and Modelling/Holdout sets

After looking through the training dataset, 8 matches contained null home and away goals entries. This is an insignificant proportion of the dataset, so can be ignored in the modelling process. Also, there is one team ID, 776, which appears in one match as an away team. This is the only team that is not in both sets of IDs.

Feature Selection

Feature engineering

Features used in the model fall into 3 main categories:

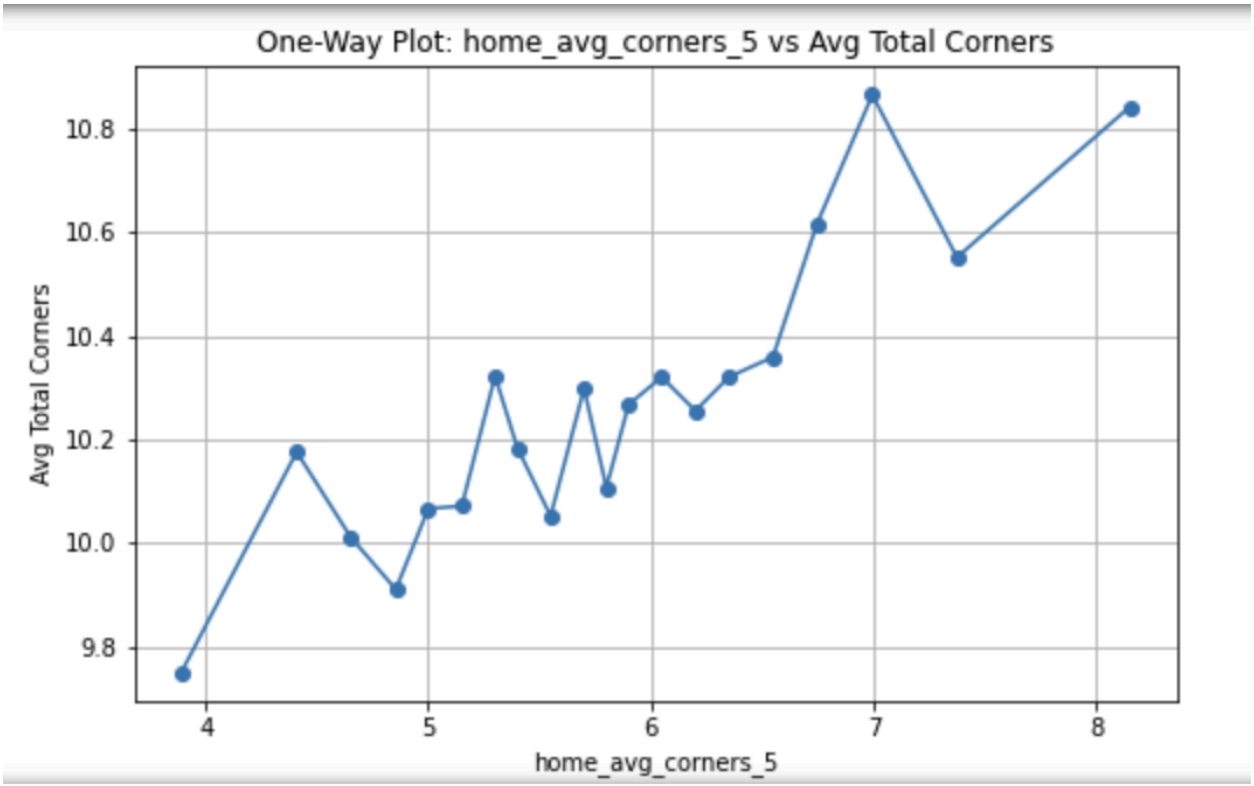
1. Goals and Corners - rolling averages over the last 5 matches of goals scored, goals conceded, corners, goal difference and corner difference each for home and away
2. Form - over the last 10 matchs, number of wins, draws, losses, points. Also points standard deviation to see if volatile performances vs consistent performance shows a trend, win difference (number of wins - number of losses - number of draws), loss difference.
3. League features - rolling averages over last 50 matches of number of corners, home corners, away corners, total goals, home goals, away goals, corner difference and goal difference.

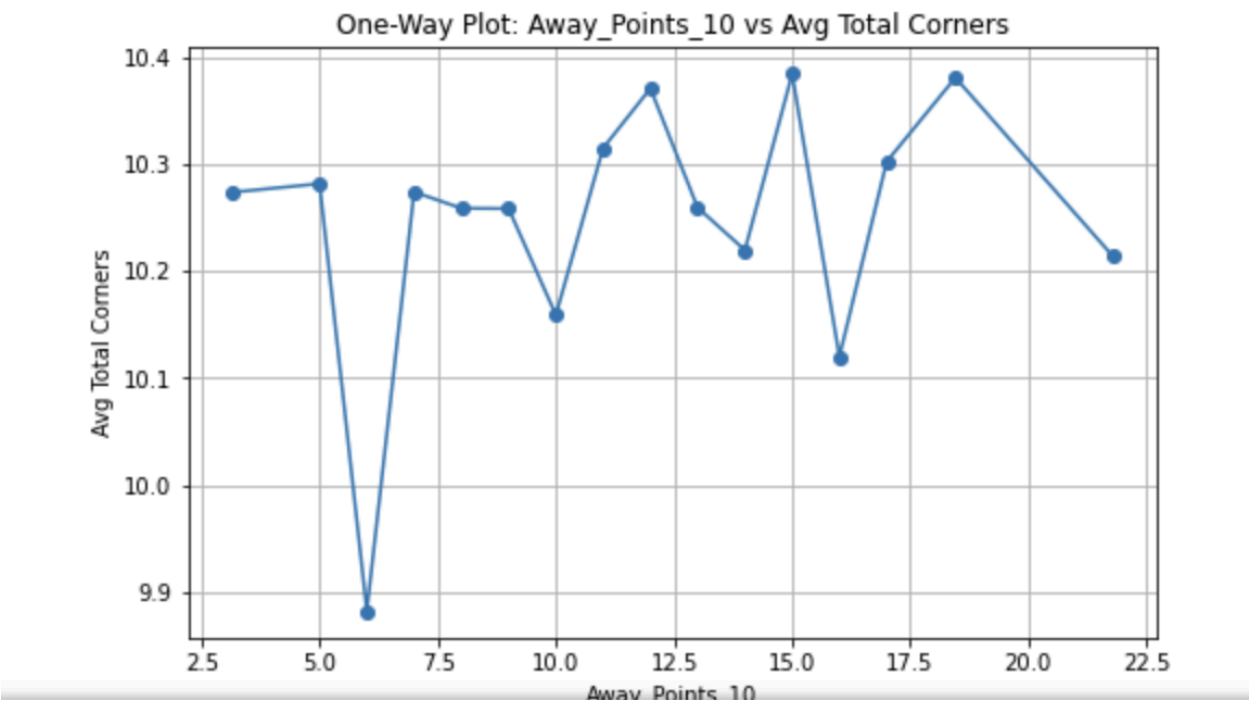
Lastly some ratios to measure teams relative to the league were created, these were corner and goals ratios, for each home and away. Eg. Home to league corner ratio = average home corners over last 5 matchs/ average league corners over last 50 matches.

Category 1 intuitively should pick up current strategy/playing style of teams, hence the more recent period selected. Category 2 should pick up team form, which I think typically changes slower than strategy, hence the longer period of 10. Lastly the league features can allow for more matches while staying relatively up to date, hence the period of 50.

One Way Charts

One way charts helped show the trend with the response and if the factor should be fitted initially. Eg. The two charts below show examples of one feature that shows a predictive trend and one that doesnt.





Interactions

A few interactions were investigated, with a list of ideas below. Due to lack of time only 3 were fitted.

1. Home goals scored : Away goals conceded - checked, next one provided lower AICcv than both together
2. Away goals scored : Home goals conceded - fitted
3. Home avg corners : home league corner ratio - fitted
4. Home league corner ratio : Away league corner ratio - fitted
5. Home avg corner/goal diff : Away avg corner/goal diff - may help identify when very dominating team plays a more passive/weaker team
6. Avg goals scored : league avg goals scored - like number 3, high scoring team in generally low scoring league could mean more attacking play in recent matches

7. Home/away Avg corner difference : avg total corners - could highlight where passive games or end-to-end comes are more likely to occur

Model Validation

A vs E Charts

Chart 1- AvE of model 3 on modelling set. This is as expected, predictions fitting closely to the observed with slight gapping where data is low in volume.

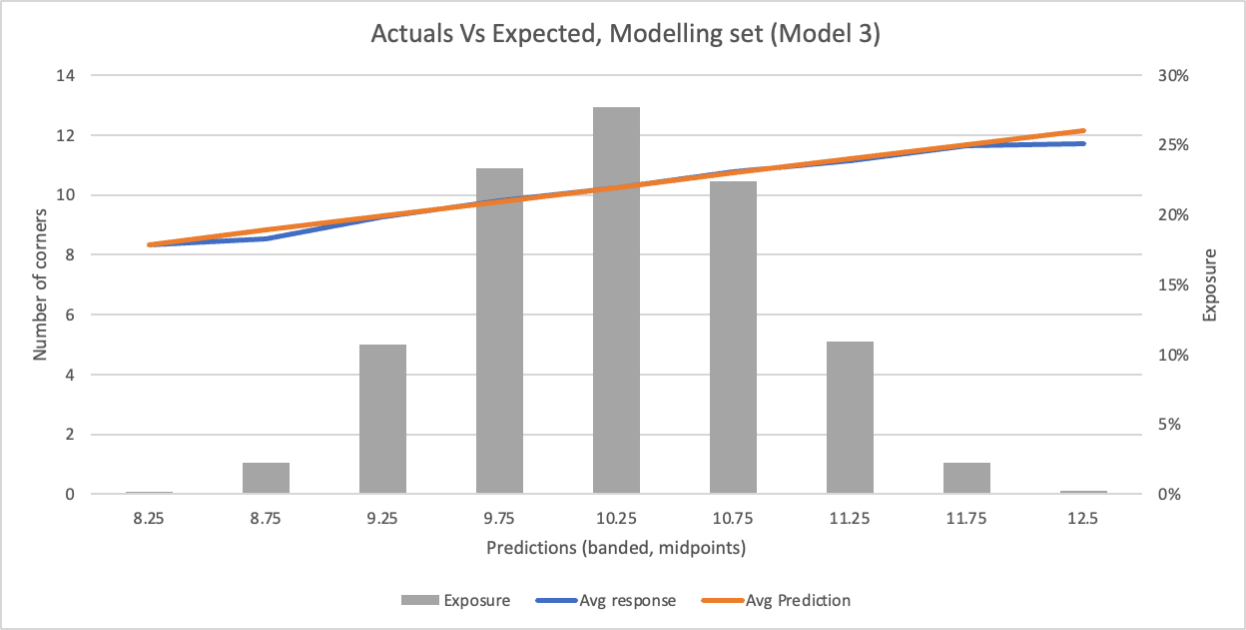


Chart 2 - AvE for Model 3 on holdout data. This shows that the model segments lower and higher corner games relatively well for the vast majority of the holdout data, with a larger gap where the data becomes low in volume. The close fit through the mass shows that the model generalises well on unseen data and hence overfitting has been well mitigated.

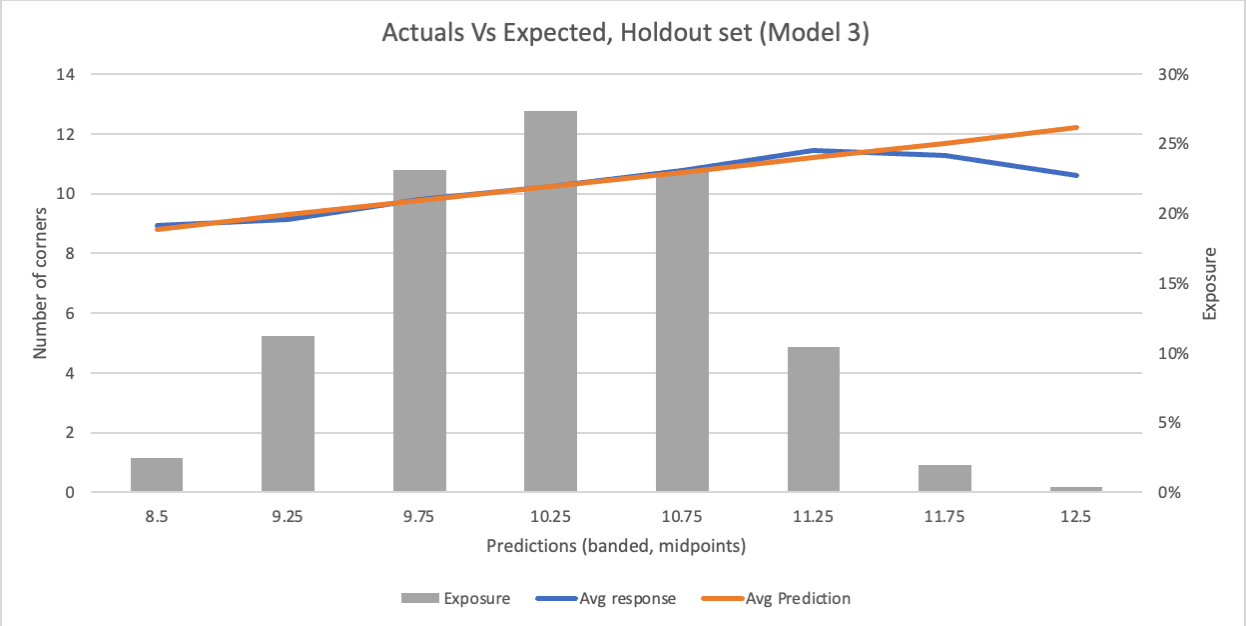
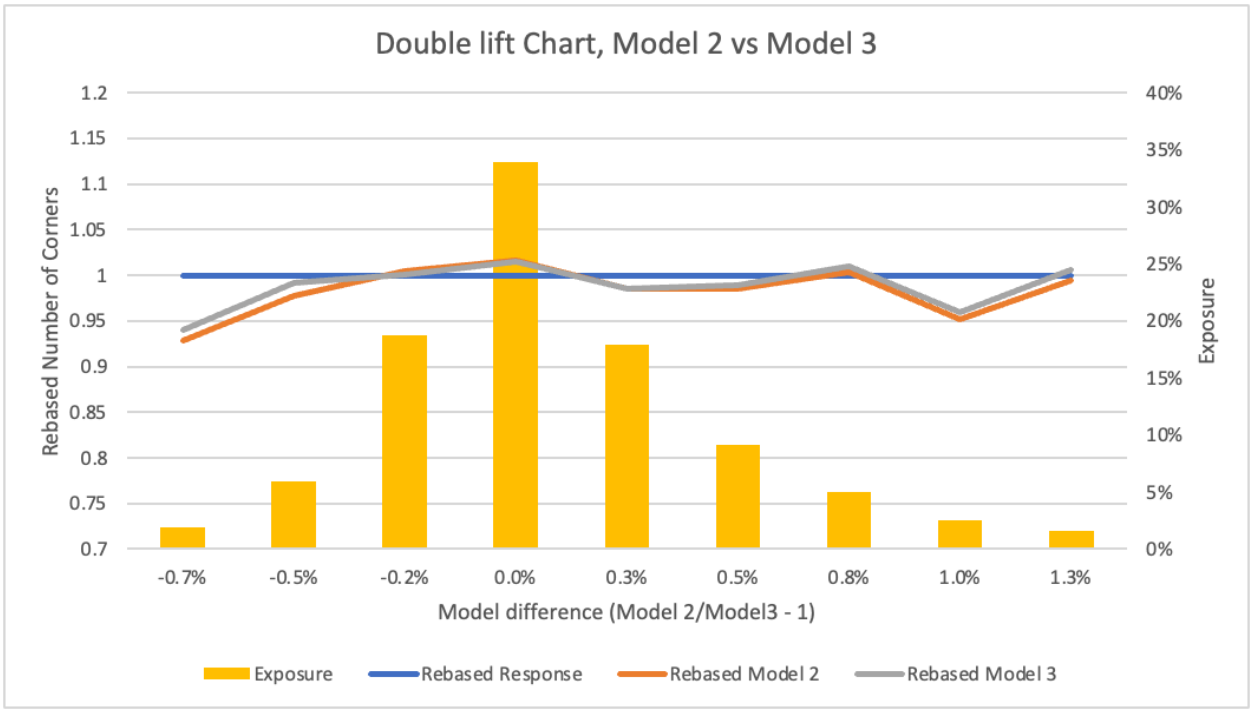

Actuals Vs Expected, Holdout set (Model 3)
Number of corners
Exposure
14
12
10
8
6
4
2
0
30%
25%
20%
15%
10%
5%
0%
8.5
9.25
9.75
10.25
10.75
11.25
11.75
12.5
Predictions (banded, midpoints)
Exposure
Avg response
Avg Prediction

Chart 3 - Double lift chart Model 2 vs Model 3 on holdout data. To recap:

Model 2 - factors showing flat trends with one-way and showing lower AICc when removed have been removed. No interactions

Model 3 - Model 2 plus the 3 interations listed in the factor list pdf.

The grey line fits closer to the blue observed indicating that Model 3 performs better on the unseen data, especially at the extremes. The models are fairly similar in performance through the middle bands, representing the mass of the data and the total model difference range is quite small, as is expected since the major difference is the inclusion of interactions.



Betting allocation

Implied Probabilities

Using the predictions of the model on the as the mean for each row, the dispersion parameter was estimated in the preliminary analysis empirically using the fact that Var = mean + mean^2/dispersion. Therefore:

Dispersion = (var - mean)/mean^2.

Now for NB(k; r,p) we can find for each row:

r=1/dispersion, p=prediction/(prediction+r).

Now for non-integer lines,

P(Under) = P(Corners <= Line - 0.5)

P(Over) = 1- P(Under)

For integer lines:

P(Under) = P(Corners <= Line -1)

P(At) = P(Corners = Line)

P(Over) = 1- P(Corners <= Line)

Optimal stake

To find the optimal stake for one bet, define for initial wealth 1

x = amount of initial wealth to bet

p = Probability of winning

q = Probability of losing

O = odds of winning

Therefore return upon winning is (O-1)*x, so the expected wealth after the bet is

$$W = (1 + (O - 1)x)^p(1 - x)^q$$

since for one bet we win p times and lose q times (or for N bets we win N*p times). Now the optimal value of x is when the below is 0:

$$\frac{d}{dx}log(W) = \frac{p}{1+(O-1)x} - \frac{q}{1-x}$$

So we get after rearranging

$$x = \frac{p(O-1)-q}{(O-1)(p+q)}$$

where p+q = 1 in the case of a non-integer line. In the case of an integer line, suppose the selected bet is to bet over. Then p is P(Over) and q is P(Under).

For the test set, there are now 341 optimal stakes for initial wealth 1, so the final stakes are for match i:

$$x^{final}_{\ i} = \frac{341x_i}{\sum_i x_i}$$